2020

Essays On Econometrics With Latent Heterogeneity And Production Function Estimation

Peng Shao
University of Pennsylvania

Essays On Econometrics With Latent Heterogeneity And Production Function Estimation

Abstract

Modeling heterogeneity among firms, workers, or countries ensures robust empirical analysis and also provides a measurement of important, but latent, economic forces. However, to model with economic theory alone is challenging and leaves room for data-driven econometric methods to assist. Furthermore, in contrast to a single cross-section, panel data offer more observations to model heterogeneity flexibly, to sharpen estimates' precision, and to reduce estimates' finite-sample bias. This dissertation develops data-driven econometric panel methods to account for heterogeneity and applies them to estimate the firm's production function for policy analysis. And there are three substantive chapters. The second chapter proposes an estimator for the partially linear model with additive time-varying grouped fixed effects. Popular empirical methods, such as regression discontinuity and control functions, employ the partially linear model. The addition of grouped fixed effects allows time-varying heterogeneity, which is a natural extension in the panel environment. The second chapter's estimator combines the use of the series approximation and the K-mean algorithm. Furthermore, I provide sufficient conditions to characterise the estimator's asymptotic performance - under large N and T but with N as comparatively larger than T. The third chapter studies an Olley-Pakes type (proxy variable) estimator for the firm's production function. Empirical economic literature extensively uses the proxy variable approach to answer policy questions on the economy's productivity and the market's competitiveness. By using the partially linear model with Grouped Fixed Effects, the third chapter extends the proxy variable approach to allow differences in firms' productivity dynamics by finitely many groups. The extension provides a new framework for empirical research to identify intrinsic differences in firms' technologies and study the inequalities among firms' performances. Using Chilean manufacturing data, I find the productivity dynamics groups explain the firm's performance in market share and the ability to export. The fourth chapter, co-authored with Xu Cheng and Frank Schorfheide, studies multidimensional latent heterogeneity in a GMM framework. We present a generalised K-mean algorithm to account for multidimensional heterogeneity in our nonlinear GMM framework. Similarly, we provide sufficient conditions to characterise the estimator's asymptotic performance - under large N and T but with N as comparatively larger than T. For application, we consider the dynamic panel estimation of the firm's production function. Here, the firms have latent heterogeneity in their output elasticities and mean productivity levels. The fourth chapter concludes in applying our estimator to document the rise of aggregate mark-up in the US economy.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Economics

First Advisor

Frank Schorfheide

Second Advisor

Xu Cheng

ESSAYS ON ECONOMETRICS WITH LATENT HETEROGENEITY AND PRODUCTION FUNCTION ESTIMATION

Peng Shao

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Co-Supervisor of Dissertation

Xu Cheng
Associate Professor of Economics

Co-Supervisor of Dissertation

Frank Schorfheide
Professor of Economics

Graduate Group Chairperson

Jesus Fernandez-Villaverde
Professor of Economics

Dissertation Committee

Amit Gandhi, *Professor of Economics*

ESSAYS ON ECONOMETRICS WITH LATENT HETEROGENEITY AND PRODUCTION FUNCTION ESTIMATION

Dedicated to my wife and my parents.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my advisors, Xu Cheng and Frank Schorfheide, and my committee member, Amit Gandhi, for their invaluable guidance and support for this dissertation to be possible. It is not possible to paraphrase here the extent of their help to do justice, but I still wish to say a few words.

I want to thank Xu for showing her unwavering faith in me to succeed and providing opportunities to do so. I want to thank Frank for teaching his exceptional standards to do research and be readily accessible to help, especially during my job market period. Moreover, I have learnt tremendously from Xu and Frank by working in our co-authored paper, presented in the fourth chapter. I am indebted to Xu and Frank for my training to research in econometrics.

Furthermore, I am grateful for Amit sharing his deep insights for his work and providing feedback for my work. I am indebted to Amit for his support and guidance. Finally, I hope this dissertation is worthy of the effort and time invested by Xu, Frank, and Amit.

There are many other professors to whom I want to express my gratitude. Francis X. Diebold trusted in me as his Econ 104 teaching assistant for so many semesters, and I wish to thank him for sharing his knowledge to design exceptional course material and experience to be an instructor with a passion for his subject. I also thank Karun Adusumilli, Frank DiTraglia, Wayne Gao, and Yuan Liao for their valuable comments in my presentation and work. Lastly, I thank David Rivers for providing me the data set used in my empirical work for the third chapter.

I also want to mention my friends and fellow graduate colleagues at Penn. I am grate-

ful to have Ashwin Kambhampati and Carlos Segura-Rodriguez as my co-authors. They show exceptional grit as we revise and polish our applied theory paper. I also wish to thank my fellow graduate colleagues in the Econometric Lunch group for their comments. Finally, I thank Paolo for being a great friend to talk with and learn from. I wish them all the best for their future endeavors.

Lastly, I wish to thank my wife for her trust and support in me. I am blessed to have a special someone during my journey to complete my dissertation.

ABSTRACT

ESSAYS ON ECONOMETRICS WITH LATENT HETEROGENEITY AND PRODUCTION FUNCTION ESTIMATION

Peng Shao

Xu Cheng

Frank Schorfheide

Modeling heterogeneity among firms, workers, or countries ensures robust empirical analysis and also provides a measurement of important, but latent, economic forces. However, to model with economic theory alone is challenging and leaves room for data-driven econometric methods to assist. Furthermore, in contrast to a single cross-section, panel data offer more observations to model heterogeneity flexibly, to sharpen estimates' precision, and to reduce estimates' finite-sample bias. This dissertation develops data-driven econometric panel methods to account for heterogeneity and applies them to estimate the firm's production function for policy analysis. And there are three substantive chapters. The second chapter proposes an estimator for the partially linear model with additive time-varying grouped fixed effects. Popular empirical methods, such as regression discontinuity and control functions, employ the partially linear model. The addition of grouped fixed effects allows time-varying heterogeneity, which is a natural extension in the panel environment. The second chapter's estimator combines the use of the series approximation and the K-mean algorithm. Furthermore, I provide sufficient conditions to characterise the estimator's asymptotic performance - under large N and T but with N as comparatively larger than T. The third chapter studies an Olley-Pakes type (proxy variable) estimator

for the firm's production function. Empirical economic literature extensively uses the proxy variable approach to answer policy questions on the economy's productivity and the market's competitiveness. By using the partially linear model with Grouped Fixed Effects, the third chapter extends the proxy variable approach to allow differences in firms' productivity dynamics by finitely many groups. The extension provides a new framework for empirical research to identify intrinsic differences in firms' technologies and study the inequalities among firms' performances. Using Chilean manufacturing data, I find the productivity dynamics groups explain the firm's performance in market share and the ability to export. The fourth chapter, co-authored with Xu Cheng and Frank Schorfheide, studies multidimensional latent heterogeneity in a GMM framework. We present a generalised K-mean algorithm to account for multidimensional heterogeneity in our nonlinear GMM framework. Similarly, we provide sufficient conditions to characterise the estimator's asymptotic performance - under large N and T but with N as comparatively larger than T. For application, we consider the dynamic panel estimation of the firm's production function. Here, the firms have latent heterogeneity in their output elasticities and mean productivity levels. The fourth chapter concludes in applying our estimator to document the rise of aggregate mark-up in the US economy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1

Introduction

Modeling heterogeneity among firms, workers, or countries ensures robust empirical analysis and also provides a measurement of important, but latent, economic forces. However, to model with economic theory alone is challenging and leaves room for data-driven econometric methods to assist. Furthermore, in contrast to a single-cross-section, panel data offer more observations to account for heterogeneity flexibly, to sharpen estimates' precision, and reduce estimates' finite-sample bias. This dissertation develops data-driven econometric panel methods to account for heterogeneity and applies them to estimate the firm's production function for policy analysis. And there are three substantive chapters.

It is instructive to motivate the econometric problem by discussing the running economic example. Economic analysis of policy questions frequently hinge on measurements of firms' performance in the economy. When micro-level firm data is available, empirical literature often models firms as production functions. So estimating production functions helps us to measure firms' performance for policy analysis.

A sound estimation strategy requires a parsimonious but flexible design on how production functions are heterogeneous in parameters, e.g., output elasticities in a Cobb-Douglas setup. For example, the design should allow the difference in firms' pro-

ductivity dynamics but not at the expense of significant loss in estimates' precision. Though economic theory suggests matching similar firms to share parameters, designing a reliable rule of matching is a non-trivial problem in practice. This dissertation applies its advanced methods to construct an automated solution by using data-driven algorithms.

Here, the estimation's theme is to model latent heterogeneity by employing unobserved group structures. Recent econometric literature has a growing interest in this modeling approach. And the dissertation contributes to this literature by proposing two novel estimators. Their implementation and asymptotic theory are in chapter two and chapter four. The chapters consider the asymptotic analysis under the experiment of large N and T, but with N as comparably larger than T. Furthermore, Monte Caro simulation suggests the estimators have excellent finite sample performance for production function estimation - even when T is small.

The dissertation also contributes to the production function estimation literature. We apply our novel estimators to extend the two conventional methods: proxy variable and dynamic panel. As mentioned in the beginning, our extensions provide a data-driven and automated process to decide which firms share production function parameters. Furthermore, our extensions are mechanically straightforward and easy to understand for empirical researchers accustomed to conventional approaches. Effectively, the extensions are both accessible and useful for empirical research.

The second chapter, "Semiparametric Panel Model with Group Heterogeneity", proposes an estimator for the partially linear model with additive time-varying Grouped Fixed Effects. The partially linear model nests a broad class of empirical economic models, such as some control function analysis and regression discontinuity. In the panel environment, economics agents' heterogeneity may evolve differently over time.

And with the addition of Grouped Fixed Effects, the partially linear model now permits different time trends.

The third chapter, "Production Function Estimation with Heterogeneous Groups", proposes a production function estimator for firms having heterogeneous productivity dynamics. I build the estimator by embedding the second chapter's partially linear model into the proxy variable estimation framework in a mechanically straightforward fashion. Furthermore, the extension's interpretation and identification follow from the widely used structural value-added model in the literature. The chapter offers a convenient solution to model heterogeneous productivity dynamics for empirical researchers accustomed to the proxy variable approach. To illustrate the extension's use, I show how the newly identified heterogeneity explains firms' ability to export and to compete for market shares in the Chilean manufacturing data set.

The fourth chapter, "Clustering for Multidimensional Heterogeneity", is co-authored with Xu Cheng and Frank Schorfheide and proposes a generalised K-mean clustering algorithm for our nonlinear GMM estimator to account for multidimensional latent heterogeneity. Here, we present a computationally straightforward extension of the dynamic panel estimation of the production function. Our proposed estimator accounts for firms' latent heterogeneity in both output elasticities and the mean productivity level. In our application, we measure the US economy's aggregate markup to revisit the question of rising market power in the US economy.

Chapter 2

Partially Linear Model with Group Heterogeneity

2.1. Introduction

Econometrics has a broad interest in modeling latent heterogeneity. Countries, firms, and workers make choices based on their rich information sets. However, empirical research widely accepts some vital information as unobserved from data and treat them as latent heterogeneity. This chapter models heterogeneity to produce robust estimation and to study unobserved economic forces.

When panel data is available, modeling latent heterogeneity as the individual fixed effect is popular but has its drawbacks. The fixed effect is time-invariant and usually involves estimating numerous parameters - placing a cost on estimates' precision. A simplified adjustment to introduce time-varying heterogeneity is modeling time intercepts shared by all. However, in many applications, time-trends are also latent heterogeneous. For example, firms have different productivity trends based on their technologies, and countries have different growth trends based on their institutions. And the quality of neither the firm's technology nor the country's institution may be observed in the data.

Here, the chapter follows an alternative model of heterogeneity, which is to assume a latent group structure, where grouped economic agents share the individual effects.

This modeling approach accounts for heterogeneous time-trends and doesn't require prior knowledge of group memberships - instead, a clustering algorithm can estimate groups from data. Many panel data settings have more cross-sectional observations than periods, and, at there, the clustering algorithm offers sharper precision for its estimates, as compared to fixed effects. In turn, the latent group structure and sharper estimates of heterogeneity also can also serve as a medium for economic analysis.

Indeed, recent literature has taken an interest in this approach. Bonhomme and Manresa (2015) grouped time-trends, with K-mean clustering, in the linear model and coined it as grouped fixed effects. More recently, Gu and Volgushev (2019) apply the grouped fixed effects to the quantile regression. Furthermore, modeling heterogeneity as cluster-specific relates to the finite-mixture likelihood model - a mature technique used widely in empirical economics.

This chapter proposes grouped fixed effects to model time-varying heterogeneity for a panel partially linear model. Popular econometric methods such as regression discontinuity and control functions use the partially linear model. Furthermore, the nonparametric regression is a special case of the partially linear model. Sequential exogeneity is allowed and, hence, the model also nests the nonparametric dynamic panel regression. With economic panel data, the partially linear model has been used to analyze firms and countries across time. In the third chapter, I present a production function estimation by using this partially linear model with grouped fixed effects. By using Chilean manufacturing data, the third chapter shows the estimated group structure can explain the firm's various characteristics useful for policy analysis and to address an empirical misspecification issue in production function estimation. This chapter provides an estimator and asymptotic analysis for the partially linear model with grouped fixed effects to facilitate economic analysis in the third chapter

and other applications.

My proposed estimator of the partially linear model uses series approximation, for nonparametric estimation, while adopting K-mean to cluster the time-trends. My theoretical contribution shows my estimator to consistently estimate the clustered group memberships under an appropriate rate of series expansion. Consequently, I establish a consistency rate for the model's nonparametric estimator and asymptotical normality for the model's linear coefficients estimator. My theoretical argument benefits from Bonhomme and Manresa (2015) and the literature of series based nonparametric estimation. Furthermore, I propose a data-driven method to select the number of groups (or clusters) and provide sufficient conditions for its consistency. All asymptotic experiments consider large N and T but having N as comparably larger than T to allow short-panel environments. However, my Monte Carlo simulation suggests the estimator performs well even under a small T.

Nonparametric estimation by series has a vast theoretical literature and is a popular empirical method, because of its convenience in implementation. For example, Olley-Pakes type production function estimators often use the power series, and Imbens and Lemieux. (2007) recommend linear splines for regression discontinuity. Chen (2007) offers other empirical examples. The theoretical foundation for series estimation is to derive the sufficient rate of series expansion for consistency. Here, I briefly name a few examples. In the cross-section, Newey (1997) derives the rate for nonparametric regression while Qi (2000) does so for the partially linear model. More recently, Belloni et al. (2015) sharpen the rate for nonparametric regression, and Lee and Robinson (2016) allow cross-sectional environment. In a similar tradition, I provide a sufficient rate for series expansion for consistency but also to achieve classification for grouped time-trends. Furthermore, my derived rate is a function of the panel's

effective sample (N, T) and allows serial dependency.

Bai and Ando (2016), Bonhomme, Lamadon, and Manresa (2017), Bonhomme and Manresa (2015), Lin and Ng. (2012), and Liu et al. (2018) use K-mean to estimate the latent groups, but they don't involve nonparametric estimation. As an alternative, Gu and Volgushev (2019) and Su, Shi, and Philips (2016) consider l_1 regularization to estimate the latent groups. K-mean and l_1 regularization each have a comparative advantage. As pointed out by Wang and Su. (2019), l_1 regularization requires to set an additional tuning parameter while K-mean involves more computation because of its local optimization nature. However, advances in modern computing with parallel operation helps to mitigate the computation burden, and the use of local optimization is familiar to many areas of empirical economics, such as the use of EM algorithm and solving Euler equations.

Recently, Wang and Su. (2019) proposes the SBSA classifier as an alternative to address the disadvantage of l_1 regularization and K-mean. However, the SBSA's idea requires every parameter as consistently estimable based on a single time-series. This idea can not apply to time-varying grouped fixed effects because there are as many parameters as the number of periods in a single time-series. Finally, Finite-Mixture approach is another alternative based on likelihood estimation. In comparison, K-mean classification is computationally convenient because it avoids the needs to correctly specify and to consistently estimate the likelihood density.

Finally, the chapter also talks to the literature on modeling heterogeneity in partially linear models and nonparametric regressions. Ai, You, and Zhou. (2014) looks at the partially linear panel model with fixed effects, under strict exogeneity, and Lee (2014) covers nonparametric dynamic panel regression with fixed effects. First, the grouped fixed effect is time-varying, whereas the fixed effect is time-invariant. Second,

under sequential exogeneity, the fixed effect model has the incidental parameter bias (Nickell (1981)) in a short panel. Indeed, Lee (2014) derives this bias as vanishing only with a larger T, and consequently, the number of series terms is subjected to T. In contrast, grouped fixed effects does not have an incidental parameter bias when N is comparably larger than T. My derived rates can allow greater freedom in choosing the number of series terms in the short panel.

The second established modeling approach is to use interactive fixed effects. To the best of my knowledge, the literature has only considered a strong factor setup in the nonparametric setting. While interactive fixed effects allow individual effects, but they are a linear combination of global factors. While grouped fixed effects restrict group members to have the same time effect, the model permits time-variation in heterogeneity to happen at a local scale. Furthermore, interactive fixed effects also have an incidental parameter bias problem in a short panel. Indeed, Huang (2013) and Su and Jin (2012) show that nonparametric regression with interactive fixed effects has a bias as vanishing only with a larger T. More recently, Freyberger (2018) considers a nonseparable generalisation of a semiparametric regression with interactive fixed effects.

The rest of the chapter is organised as follows. In Section 2.2, I set up the partially linear model with its three motivating examples. Then I describe my proposed estimator in Section 2.3 and establish the asymptotic analysis in Section 2.4. In Section 2.5, I present a Monte Carlo simulation for finite sample performance. The conclusion happens in Section 2.6, then follows by extensions in Section 2.7.

2.2. Model and Examples

The partially linear semiparametric panel model is,

$$y_{it} = x'_{it}\theta^0 + m\left(z_{it}\right) + \alpha^0_{it} + \epsilon_{it}, i, = 1, ..., N, t = 1, ..., T, \tag{2.1}$$

where the variables $(y_{it}, x_{it}, z_{it}, \alpha_{it}, \epsilon_{it}) \in \mathbb{R} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{A} \times \mathbb{R}$ ($\mathcal{Z} \subset \mathbb{R}^{d_1}$, $\mathcal{X} \subset \mathbb{R}^{d_2}$, and $\mathcal{A} \subset \mathbb{R}$), the unknown function $m : \mathcal{Z} \to \mathbb{R}$, and the parameter $\theta^0 \in \Theta$, with $\Theta \subset \mathbb{R}^{d_2}$.

There are G^0 fixed groups, and each unit belongs to a group. Furthermore, each unit has an unobserved and time-invariant membership with index $g^0_i (\in \Gamma_{G^0} := \{1, ..., G^0\})$. Within a group g, all its members share the time trajectory α_{gt}.

So,

$$\alpha^0_{it} = \begin{cases} \alpha^0_{1t} & g^0_i = 1 \\ \vdots & \vdots \\ \alpha^0_{G^0 t} & g^0_i = G^0. \end{cases} \tag{2.2}$$

The chapter purposes this model for two objectives. The first intends to conduct inference on θ when m and α_{it} act as nuisances. The second studies its as a non-parametric regression, i.e. $\theta^0 = 0$, and aims to estimate the conditional mean m and the heterogeneity parameter α_{it}. To fix ideas, Example 1 and Example 2 provide an economic example for each objective, respectively. Therefore, the main interests are to provide an estimator for inference on θ and consistent estimators for (m, α^0_{it}).

I assume the econometrician observes the outcome variable y_{it} and the regressors (x_{it}, z_{it}). Furthermore, I allow the regressors (x_{it}, z_{it}) as endogenous of $\alpha^0_{g^0_i t}$ but sequentially exogenous of ϵ_{it}. This setup is general enough to nest the dynamic panel

model. For identification, I normalise $m(z') = 0$ for some $z' \in \mathcal{Z}$ because of treating m as nonparametric.

Example 1: (Income Growth vs. Inequality Growth) Banerjee and Duflo (2003) studies whether an increase in income inequality overall promotes or hinders, $\left(\theta_1^0\right)$, income growth, $\Delta income_{it}$, in a panel of countries. In their equation (9), income growth is affected by both inequality growth, $\Delta Gini_{it}$, and level, $Gini$. However, only inequality growth is assumed to have a linear effect, while the inequality level may have a non-linear effect. Here, I present a simplified version as,

$$\Delta income_{it} = \theta_1^0 \Delta Gini_{it} + \theta_2^0 \Delta income_{it-1} + m\left(Gini_{it}\right) + \alpha^0_{g_i^0 t} + \epsilon_{it}. \tag{2.3}$$

Different economic policies and political institutions have varying effects on income growth. In their setup, a time-invariant country fixed effects captures the net outcome of these effects. Here, I propose to model these net effects as time-varying grouped fixed effects, $\alpha_{g_i^0 t}$, because policies' effects are time-varying. The groups partition countries by their independence of the judiciary, and market-based vs. central planning spectrum. It is reasonable to allow inequality growth as endogenous of inequality level and institutions. So to do correct inference on θ_1^0 requires to control for m and α_{it}.

Example 2: (Production Function estimator) The general intent is to estimate Hicksian Neutral production functions but, for simplicity, I illustrate the problem with a log-linearized Cobb-Douglas production function. The aim is to introduce heterogeneity in productivity dynamics within the Ackerberg, Caves, and Frazer (2015)'s framework. Similar to their first-stage estimation, here there is a nonparametric

regression as specified by

$$y_{it} = m\left(l_{it}, k_{it}, \upsilon_{it}\right) + \alpha_{g_i^0 t} + \epsilon_{it}, \tag{2.4}$$

where y_{it}, l_{it}, k_{it}, and υ_{it} are logged output, logged labour, logged capital, and logged material inputs, respectively. ϵ_{it} is a mean-zero idiosyncratic productivity shock, happening after the firm's input decisions. The example adds a new productivity term $\alpha_{g_i^0 t}$ for firms to have heterogeneous productivity dynamics. The simplest example is to model firms belonging to two groups within an industry. Then the first group of firms shares a higher productivity persistence over the second group of firms'. Alternatively, the differences between the two groups can be in mean-levels and/or, more generally, the functional form of productivity transition, e.g., moving average vs. autoregressive. These differences are plausible in application because an industry consists of firms with varying technologies in production and management. All these observations easily extend to a model of multiple groups.

Let β_l and β_k be the output elasticities of labour and capital inputs. Furthermore, let ω_{it} be a firm-specific productivity component. Chapter 3 shows $m\left(l_{it}, k_{it}, \upsilon_{it}\right) = \beta_l l_{it} + \beta_k k_{it} + \omega_{it}$ under a Cobb-Douglas specification in a modified structural value-added model - see Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2017a) for examples of structural value-added models in the production function literature.

For now, I assume ω_{it} as a mean-zero and independent over time to simplify presentation - Chapter 3 generalises ω_{it} to be a first-order Markov process. The independence assumption implies the firm's expectation of ω_{it} as zero at period $t-1$.

Hence, the following moment conditions identify the output elasticities, (β_k, β_l),

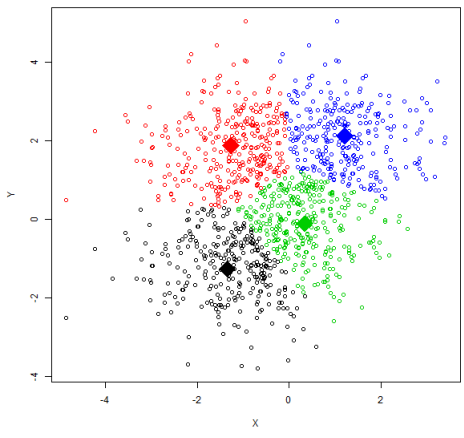
$$\mathbb{E}\left[\begin{pmatrix} k_{it-1} \\ l_{it-1} \\ 1 \end{pmatrix} \left(m\left(k_{it}, l_{it}, \upsilon_{it}\right) - \beta_k k_{it} - \beta_l l_{it} - \nu\right)\right] = 0, \tag{2.5}$$

for some constant ν. Introducing the constant ν accounts for m's normalisation. Conditional on having m then a moment based estimator for output elasticities is available. This observation alludes to construct a two-step estimator. The first-stage estimates m and, subsequently, the second-stage estimates the output elasticities by using the estimated m. This feasible estimator is consistent when the first-stage consistently estimates m. Chapter 3 discusses the asymptotic in more detail.

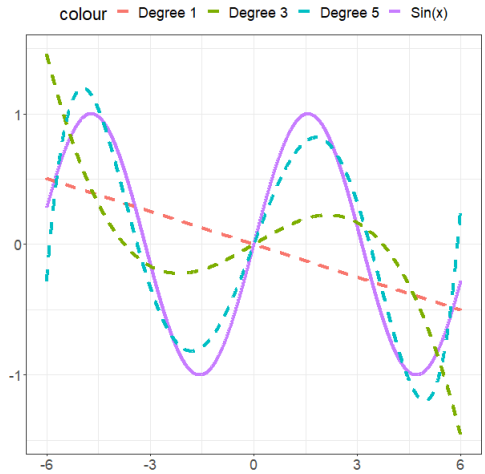
2.3. Estimation

Estimation

To estimate the parameters $\left(\theta^0, m, \alpha_{gt}, g_i^0\right)$, I simultaneously apply two econometric techniques: K-mean clustering to classify g_i^0 and series approximation to estimate m.



K-Mean Clustering



Series Approximation: Power Series

The principle of K-mean is to detect the group structure, g_i^0, by partitioning the data around centroids. The left graph shows K-mean in action. Dotted observations are classified with color around four centroids, marked by diamond shapes. In the partially linear model's context, the centroids can be interpreted as the parameters, and the dotted observation's distance from its centroid is the residual ϵ_{it}.

The principle of series approximation is to use the sum of smooth functions to approximate an unknown function, m. The right graph shows this principle in action. The polynomials of x are trying to approximate $Sin\left(x\right)$, and the approximation error vanishes by increasing the polynomial degree. For smooth function m, the series approximation can be thought of as m's Taylor approximation.

Next I describe how to implement the two procedures jointly. Suppose the econometrician assumes the number of group is G and estimates the non-parametric $m(\cdot)$ with a vector of basis functions, $p^K(\cdot) = \left(p_1\left(\cdot\right), ..., p_K\left(\cdot\right)\right)$, where $p_s\left(\cdot\right) : \mathcal{Z} \to \mathbb{R}$, for $s = 1, ..., K$, and K is an integer. Furthermore, $\beta^K = \left(\beta_{1K}, ..., \beta_{KK}\right)'$ is the vector of coefficients for $p^K\left(z_{it}\right)$ to approximate m and $\beta_{sK} \in \mathcal{B}^K$. Popular example of p^K includes power series, Fourier series, and B-splines.

The group assignment $\gamma : \{1, ..., N\} \to \Gamma_G$, where $\gamma\left(i\right) = g_i$ and $\Gamma_G := \{1, ..., G\}$, denotes the collection of group membership parameters - with $\gamma^0 := \{g_i^0\}_{i=1}^N$. The partially linear model's estimator comes from minimizing the least-squared criterion:

$$\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) \in \underset{\theta \in \Theta, \beta^K \in \mathcal{B}^K, \alpha \in \mathcal{A}^{G \times T}, \gamma \in \Gamma_G^N}{\arg\min} \hat{Q}(\theta, \beta^K, \alpha, \gamma), \tag{2.6}$$

where $\hat{Q}\left(\theta, \beta^K, \alpha, \gamma\right) = \dfrac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta^K - p^K\left(z_{it}\right)'\beta^K - \alpha_{g_i t}\right)^2$. The estimator of m is $\hat{m}(z_{it}) = p^K(z_{it})'\hat{\beta}^K$.

The least-squared minimization problem is non-linear in γ. Next, I provide a simple local optimization algorithm to solve the problem. The optimization is local because it has the least-squared solution as a convergent point but its set of convergent points may not be singular.

Algorithm 1: Estimating θ, β^K, g_i, and α_{gt}

Initialize $\left\{\hat{g}_{i[0]}\right\}_{i=1}^{N}$.;

Using $\left\{\hat{g}_{i[0]}\right\}_{i=1}^{N}$, estimate $\hat{\theta}_{[0]}, \hat{\beta}_{[0]}^{K,j}$, and $\hat{\alpha}_{gt[0]}$ by minimizing the least-squared criterion;

while *convergence is not achieved on the kth iteration* **do**

Using the kth iteration's $\hat{\theta}_{[k]}, \hat{\beta}_{[k]}^{K,j}$ and $\hat{\alpha}_{gt[k]}$, update $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$ to minimize the least squared criterion;

Using $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$, estimate $\hat{\theta}_{[k+1]}, \hat{\beta}_{[k+1]}^{K,j}$ and $\hat{\alpha}_{gt[k+1]}$ by minimizing the least-squared criterion;

Check for convergence of the modified least squared criterion;

end

The algorithm is convergent because it decreases the least-squared criterion at every step. For the global optimum, it is paramount to compute and compare local solutions using different initial arrangements of $\left\{\hat{g}_{i[0]}\right\}_{i=1}^{N}$.

Selection

In practice, the econometrician has to choose a G without knowing G^0. I assume the econometrician knows an upper bound $G_{\max}$ and a lower $G_{\min}$ for G^0, i.e. $G_{\min} \leqslant G^0 \leqslant G_{\max}$. For panel models, the information criterion is a popular tool for selecting the latent structure's complexity - see Bai and Ng (2002) and Su, Shi, and Philips

(2016). For my partially linear model, the Information Criterion function is

$$IC\left(G\right)=\hat{Q}_{G}+\nu G, \tag{2.7}$$

where $\hat{Q}_G$ is the minimized least-squared criterion from using G groups and for some positive constant ν. The information criterion estimate of G^0 is

$$\hat{G}^{0}\in\underset{G\in\{G_{\min},...,G_{\max}\}}{\arg\min}\ IC\left(G\right). \tag{2.8}$$

I provide guidance in choosing ν in the Chapter 2's Asymptotic Theory section and provide an explicit example in the Chapter 3's Monte Carlo section. The paper also provides the information criterion's consistency result.

2.4. Asymptotic Theory

This section first proves the consistency of $\hat{\theta}$ and $\hat{m}$, as presented in Theorem 1. Theorem 2 proves the consistency of $\hat{G}$ when the information criterion's penalty satisfies appropriate conditions. Then Theorem 3 shows classification error disappears asymptotically, i.e. $\hat{g}_i$ is uniformly consistent over i. Finally, Theorem 4 proves the asymptotic normality of $\hat{\theta}$ and uniform consistency of $\hat{\alpha}_{\hat{g}_i t}$.

The paper assumes the series $p^K\left(z_{it}\right)$ to satisfy some high-level properties. For interpretation, high-level assumptions are elaborated for the power series and the B-spline series. The standard theory for the power series and the B-splines assume $\mathcal{Z}$ as a compact support. So the high-level assumptions are discussed with examples in the context of $\mathcal{Z}$ being compact. Chen (2007) provides other examples of series for the compact support. The discussion of the power series and B-splines can also apply to those series. All the assumptions are sequentially presented before theorems and

progressively stronger to derive more demanding results. The provided asymptotic theory considers $N, T, K \to \infty$, but they grow at different rates. In particular, N is assumed to be significantly larger than T. Appendix-Chapter 2 collects all the asymptotic results.

Notation: Let $\|\cdot\|$ be the Euclidean norm, $\|f\|_{\infty,\mathcal{Z}} := \sup_{z\in\mathcal{Z}} \|f(z)\|$, $q_{it} = \begin{pmatrix} x_{it}' & p^K(z_{it})' \end{pmatrix}'$, $\alpha = \left\{\{\alpha_{gt}\}_{t=1}^{\infty}\right\}_{g=1}^{G^0}$, and $x_{it} = (x_{it,1}, ..., x_{it,d_2})'$.

Assumption 1. *(Series approximation)*

There exist a constant $\mu > 0$ and the sequence $\left\{\left(\xi_K, \beta^{0,K}, \Pi_K\right)\right\}_{K=1}^{\infty}$, $\left(\xi_K, \beta^{0,K}, \Pi_K\right) \in \mathbb{R}_+ \times \mathcal{B}^K \times \mathbb{R}_+$, such that:

1. *$\|p^K\|_{\infty,\mathcal{Z}} \leqslant \xi_K$, and $\xi_K \to \infty$, as $K \to \infty$.*

2. *$\sup_{\beta\in\mathcal{B}^K} \max_{l=1,...,K} \|\beta_{lK}\| \leqslant \Pi_K$, where $\beta = (\beta_{1K}, ..., \beta_{KK})$, and Π_K is uniformly bounded away from 0.*

3. *$\left\|m - (p^K)'\beta^{0,K}\right\|_{\infty,\mathcal{Z}} = O(K^{-\mu})$ and $\beta^{0,K} \in \mathcal{B}^K$.*

Assumption 1.1 requires every finite-termed series to be bounded over the support $\mathcal{Z}$. In the following assumptions, the bound ξ_K should increase at a certain rate in proportion to the sample size. Newey (1997) provides a ξ_K as proportional to K for power series [1] and $\sqrt{K}$ for B-splines. Under Assumption 1.3, the series' approximation error of m over the entire support $\mathcal{Z}$ vanishes as the series' terms increase. The μ parameter captures the smoothness of the m function for both power series and B-splines. Newey (1997) shows $\mu = \delta^d/d_1$, where δ^d is number of m's continuous derivatives. Analogous Assumption 1.1 and Assumption 1.3 can be found in Newey (1997). Assumption 1.2 introduces a notation on the upper bound of $\beta^{0,K}$'s magni-

[1] The proportionality comes the orthogonal polynomial; which it spans the same linear space as the power series do.

tude. The bound Π_K may increase over K subjected to the rates discussed later on. Π_K is constant for the power series if Assumption 1.3's approximation is also absolutely convergent in a neighborhood outside the unit ball. The examples include m being the sum of exponential functions, polynomials, and logarithms (when log takes values uniformly bounded away from zero). When Π_K is constant, the subsection's rates can drop the factor Π_K.

Assumption 2. *(Compactness)*

$\mathcal{A}$ and Θ are compact.

The compactness of $\mathcal{A}$ rules out nonstationary α_{gt} process with its mean level growing over time. This same restriction is assumed by Bonhomme and Manresa (2015).

Assumption 3. *(Dependency and moment restrictions)*

There exist a constant $M > 0$ such that

1. $\sup_{i\in\{1,\ldots,N\}} \mathbb{E}\left[\|x_{it}\|^4\right] \leqslant M$ *and* $\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}[\epsilon_{it}\epsilon_{is}x_{it}'x_{is}]\right| \leqslant M.$
2. $\mathbb{E}\left[u_{it}\right] = 0$ *and* $\sup_{i\in\{1,\ldots,N\}} \mathbb{E}\left[u_{it}^4\right] \leqslant M.$
3. $\left|\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} Cov(\epsilon_{it}\epsilon_{jt}, \epsilon_{is}\epsilon_{js})\right| \leqslant M.$
4. $|\mathbb{E}\left[\epsilon_{is}\epsilon_{it}|z_{is}, z_{it}\right]| \leqslant M$ *and* $\mathbb{E}\left[\epsilon_{is}\epsilon_{jt}|z_{is}, z_{jt}\right] = 0$ *for any* $i \neq j$.

Assumption 3 restricts the dependency of ϵ_{it} on (x_{is}, z_{is}). Both x_{is} and z_{is} can be predetermined regressors. Overall, similar assumptions can be found in Bonhomme and Manresa (2015) and Bai (2009). However, Assumption 3.4 does not allow unconditional cross-correlation of ϵ_{it} and it also imposes bounded conditional heteroskedasticity and serial correlation. The bounded conditional heteroskedasticity assumption is standard in the series literature. If z_{it} is strictly exogenous then cross-correlation

of ϵ_{it} can be restored by adopting Lee and Robinson (2016)'s approach.

Assumption 4. *(Rank Condition)*
Let $N^* := \lfloor \frac{N}{G} \rfloor - 1$ *and* $\mathcal{S} \subset \{1, ..., N\}$. *If* $\mathcal{S}$ *has at least* N^* *units then*

$$\mathbb{P}\left(\lambda_{\max}\left(\left[\frac{1}{TNN_S^2} \sum_{t=1}^{T} \sum_{i \in \mathcal{S}} \left(\sum_{j \in \mathcal{S}} (q_{it} - q_{jt}) \right) \left(\sum_{j \in \mathcal{S}} (q_{it} - q_{jt}) \right)' \right]^{-1} \right) < c \right) \underset{as\ N,T,K \to \infty}{\to} 1,$$

where $c > 0$, $\lambda_{\max}$ *is the largest eigenvalue, and* $N_S = \sum_{i=1}^{N} \{i \in S\}$.

Assumption 4 provides the rank condition to compute the least square estimator of $\left(\theta, \beta^K, \alpha\right)$, based on the estimated group memberships. For an arbitrary large group with at least N^* memberships, Assumption 4 requires sufficient cross-sectional variation of x_{it} and $p^K(z_{it})$ within the group. Hence, x_{it} and $p^K(z_{it})$ excludes constants.

Assumption 5. *(Rates and smoothness for consistency)*
As $N, T, K \to \infty$,

1. $K^{\frac{1}{2}-\mu} \xi_K^3 \Pi_K \to 0$.
2. $\dfrac{\xi_K^3 \sqrt{K} \Pi_K}{\sqrt{N}} \to 0$.
3. $\dfrac{\xi_K^2}{(NT)^{\frac{1}{4}}} \to 0$.

Assumption 5.1 controls effects from the vanishing approximation error. For power series and B-splines, Assumption 5.1 assumes m to be sufficiently smooth. Under the discussion of Assumption 1, when m has at least $4d_1$ continuous derivatives and Π_K is bounded, e.g., real analytic functions, then Assumption 5.1 holds for both power series and B-splines. Assumption 5.2 and 5.3 restricts the series terms to asymptotically grow at a slower rate than N and T. Assumption 5.2 controls effects from the

estimation error of the series' coefficients. However, Assumption 5.2's rate can be weakened to $\frac{\xi_K^3\sqrt{K}\amalg_K}{\sqrt{NT}} \to 0$ under the weak time dependency condition as specified in Assumption 9. In the cross-section setting, Newey (1997)'s semiparametric model and Qi (2000)'s partially linear model asks for $\frac{\xi_K\sqrt{K}}{\sqrt{N}} \to 0$. Assumption 5.2's rate is slower partly because parameters and group memberships are jointly estimated.

Assumption 5.3 is introduced to handle the unobserved group memberships. When N is comparably larger than T, these terms offer greater flexibility in choosing K as opposed to rates governed just by T.

Theorem 1. *(Consistency) Suppose Assumptions 1-5 hold and $G \geqslant G^0$, then 1) $\hat{\theta} \xrightarrow{P} \theta^0$, and 2) $\|m - \hat{m}\|_{\infty,\mathcal{Z}} \xrightarrow{P} 0$, as $N, T, K \to \infty$.*

Whenever the number of used groups is not smaller than the truth, Theorem 1 provides the consistency of $\hat{\theta}$ and the uniform consistency of $\hat{m}$. For just $\hat{\theta}$'s consistency, all Assumption 5's rates can be scaled down by ξ_K^2. However, having ξ_K^2 helps to show $\hat{m}$'s consistency. Moreover, subsequent results on classification and $\hat{\theta}$'s asymptotic normality depend on $\hat{m}$ being consistent. Improving Assumption 5's rates is an avenue for future work.

Assumption 6. *(Identifying Groups)*

1. *There exists a constant $c > 0$ such that,*

   (a) *when $g \neq g'$, then $\operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{g't}^0\right)^2 > c$, for a $c > 0$. This lower bound c applies to all pairs of g and g'.*

   (b) *for a real-valued process $\{h_t\}_{t=1}^{\infty}$ satisfying $\infty > \operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h_t^2 > \frac{1}{2}c$, then*

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h_t^2 > \operatorname*{plim}_{N,T,K\to\infty} \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} h_t q_{it}'\right)\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}q_{it}'\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}h_t\right).$$

2. *Let* $N_g = \sum_{i=1}^{N} \{g_i^0 = g\}$. *For any* $g \in \{1, ..., G^0\}$, $\frac{N_g}{N} \underset{N\to\infty}{\to} \kappa_g > 0$.

3. *Assume* $G = G^0$.

Assumption 6 provides the conditions to identify the groups. Assumption 6.1.a requires groups to be separately identified from their time-paths. Assumption 6.1.b is an identification assumption for the information criterion selection to avoid under-selecting. It basically says the differences between the groups effects should be far away from the regressors' spanned linear space. And Assumption 6.2 assumes each group's memberships is proportionally significant to the overall cross-section's sample size.

Corollary 1. *(Time-path consistency) Under Assumption 1-5, 6.1.a, 6.2, and* $G^0 \leqslant G$, *for any* $g \in \{1, ..., G^0\}$, *there exists a* $\hat{g}$ *such that* $\underset{N,T,K\to\infty}{\text{plim}} \frac{1}{T} \sum_{t=1}^{T} \left(\alpha_{gt}^0 - \hat{\alpha}_{\hat{g}t}\right)^2 = 0$. *Under Assumption 6.3,* $\hat{g}$ *is unique.*

With Corollary 1, each true group's α_{gt} time-path is matched asymptotically close to an estimated group's estimated time-path on average over time. This is weaker than uniform consistency but uniform consistency is achieved after additional stronger assumptions. Uniform consistency further provides consistency of $\hat{\alpha}_{\hat{g}_i t}$ at every i and t.

Assumption 7. *(Rates and smoothness for selection)*
As $N, T, K \to \infty$,

1. $\nu_T \to 0$.

2. $T^{\frac{1}{4}} \nu_T \to \infty$.

3. $\frac{T^{\frac{1}{4}} \xi_K \text{II}_K \sqrt{K}}{\sqrt{N}} = O_p(1)$.

4. $T^{\frac{1}{4}}K^{\frac{1}{2}-\mu}\xi_K \text{II}_K = O_p(1)$.

5. $G^0 \in \{\underline{G}, \overline{G}\}$.

Under Assumption 7.5, the information criterion minimizes over a set of specifications containing G^0. Under a large sample size, the logic behind the information criterion relies on over-specifying $G\left(> G^0\right)$ to yield negligible improvement and under-specifying $G\left(< G^0\right)$ leaves significant room for improvement in the least squared fit. To detect underfitting, Assumption 7.1 requires the penalty to vanish asymptotically. Moreover, to detect overfitting, Assumption 7.2 requires the penalty to vanish slowly at a rate dependent on only T. The T only dependency is set up under the assumption of N as comparably larger than T. In that environment, it is consistent with the information criterion literature to have the error rate as independent of N. For example, Bai and Ng (2002) have their rates as independent of N in the interactive factor setup when $\frac{\sqrt{T}}{N} \to 0$ - which is consistent with Assumption 7.3.

The information criterion's strategy for consistent selection also relies on the difference between the criteria, from over-specified and exactly specified, to vanish at a rate faster than the penalty. Assumption 7.3 and 7.4 execute this task in combination. Furthermore, Assumption 7.3[2] and 7.4 act as Assumption 5.1 and 5.2 for the selection purpose, respectively.

Theorem 2. *(Selection) Suppose Assumption 1-4, 6.1, 6.2, and 7 hold, then*

$$\lim_{N,T,K\to\infty} \mathbb{P}(\widehat{G^0} = G^0) = 1.$$

[2]Just like Assumption 5.2, Assumption 7.3 can be weakened to $\frac{T^{\frac{1}{4}}\xi_K \text{II}_K\sqrt{K}}{\sqrt{NT}} = O_p(1)$ under weak time dependency.

For the previous specific penalty, the selection is consistent for each λ. Hence, the data-driven choice of pre-specified λs is also consistent because the pre-specified set is finite; hence, the criterion is consistent for any λ of the set.

Theorem 2 shows the information criterion's estimate of G^0 is asymptotically consistent. Knowing the true number of groups is assumed for the subsequent theorems. However, the subsequent theorems do not account for the post-selection estimator.

Assumption 8. *(Tail-bounds)*

1. *There exist constants* $r_1 > 0$ *and* $r_2 > 0$ *such that,* $\mathbb{P}\left(|\epsilon_{it}| > m\right) \leqslant e^{1-\left(\frac{m}{r_1}\right)^{r_2}}$, *for all* i, t *and* $m > 0$. *For any* $i \in \{1, ..., N\}$,.

2. *For any* $g_l^0, g_k^0 \in \{1, ..., G^0\}$, $\mathbb{E}\left[\left(\alpha_{g_l^0 t}^0 - \alpha_{g_k^0 t}^0\right)\epsilon_{it}\right] = 0$.

3. *There exists constants* $r_3 > 0$ *and* $r_4 > 0$ *such that,* $\{\epsilon_{it}\}_{t=1}^{\infty}$, $\left\{\alpha_{g_j^0 t} - \alpha_{g_i^0 t}\right\}_{t=1}^{\infty}$ *and* $\left\{\left(\alpha_{g_j^0 t} - \alpha_{g_i^0 t}\right)\epsilon_{it}\right\}_{t=1}^{\infty}$ *are strongly mixing process, with mixing coefficient* $\rho_i(t)$, *and* $\sup_{i \in \{1,...,N\}} \rho_i(t) \leqslant e^{-r_3 t^{r_4}}$, *for any* $g_l^0, g_k^0 \in \{1, ..., G^0\}$.

4. *There exist constants* $M^* > 0$ *and* $\delta > 1$ *such that,*

$$\sup_{i \in \{1,...,N\}} T^{\delta} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \|x_{it}\| \geqslant M^*\right) \to 0$$

and $\dfrac{N}{T^{\delta-1}} \to 0$, *as* $T, N \to \infty$.

Assumption 8.1, 8.2, and 8.3 assume tail bounds and weak dependency to control for classification error on the group membership estimate. Assumption 8.4 holds if x_{it}'s support is compact, and N is comparable to some power of T. Besides the comparable size of N and T, Assumption 8 is near identical to Bonhomme and Manresa (2015)'s Assumption 2 for their linear model case.

Theorem 3. *(Group Consistency) Let* $H_g^0 := \{i \mid g_i^0 = g\}$ *and* $\hat{H}_g := \{i \mid \hat{g}_i = g\}$. *When* $G^0 = G$ *and Assumption 1-6 and 8 hold, for any* $g \in \{1, ..., G\}$, *there exists* $g^0 \in \{1, ..., G^0\}$ *such that* $\mathbb{P}\left(\hat{H}_g = H_{g^0}^0\right) \to 1$, *as* $N, T, K \to \infty$.

When the exact total number of true groups is used, Theorem 3 says every estimated group asymptotically match to a true group in memberships. After some re-labeling of groups, Theorem 3 implies the uniform consistency of the group membership estimate, i.e $\mathbf{P}\left(\sup_{i\in\{1,...,N\}} |\hat{g}_i - g_i^0| > 1\right) = o_p(1)$. Furthermore, Theorem 3 also leads to $\left(\hat{\theta}, \hat{m}, \hat{\alpha}\right)$ as asymptotically equivalent to the Oracle estimator $\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}_{gt}\right)$. The Oracle estimator minimizes the least-squared criterion based on the true memberships.

The next assumption provides the additional conditions leading to uniform convergence of $\hat{\alpha}_{\hat{g}_i t}$, rate for $\hat{m}$, and the asymptotic normality of $\hat{\theta}$. To simplify presentation, I assume the moments, conditional of α, are identical within group. The proof uses the extended version, allowing heterogeneous conditional moments within the group. The extended version is provided in Appendix-Chapter 2.

Assumption 9. *(Asymptotic Normality)*

1. *There exists a* $\delta_m > 0$ *such that* $\max\left\{\left\|\theta - \theta^0\right\|, \left\|\beta^K - \beta^{0,K}\right\|\right\} < \delta_m$ *implies* $\beta^K \in \mathcal{B}^K$ *and* $\theta \in \Theta$. *Furthermore,* $\hat{\alpha}_{gt}$ *is the interior solution.*

2. *For each* $i \in \{1, ..., N\}$,

   (a) *Conditional on* α, $\{(x_{it}, z_{it}, \epsilon_{it})\}_{t=1}^{\infty}$ *is independent over* i.

   (b) *Both conditional and unconditional on* α, $(x_{it}, z_{it}, \epsilon_{it})$*'s alpha mixing coefficient satisfies the uniform bound described in Assumption 8.3 up to a scale.*[3]

[3]It is not necessarily to assume Assumption 9.2 shares the same parametric values of r_3 and r_4

(c) ϵ_{it} *is mean independent of* x_{it}, z_{it}, *and* α.

(d) *the* $(x_{it}, z_{it}, \epsilon_{it})$ *process is stationary.*

3. *With some constant* $C^{xp} > 0$, *for any fixed* K *and* $g \in \{1, ..., G^0\}$, *the matrix*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(q_{it} - \mathbb{E}\left[q_{it}|\alpha\right]\right)\left(q_{it} - \mathbb{E}\left[q_{it}|\alpha\right]\right)'|\alpha, g_i^0 = g\right]' s$$

smallest eigenvalue is bounded below by C^{xp}.

4. *Let* $v\left(z_{it}\right) := \left(\mathbb{E}\left[x_{it,1}|z_{it}\right], ..., \mathbb{E}\left[x_{it,d_2}|z_{it}\right]\right)$.

(a) *There exist sequences* $\left\{\left\{\beta_{x,j}^K\right\}_{k=1}^{\infty}\right\}_{j=1,...,d_2}$ *and constants* $\{c_{vj}\}_{j=1,...,d_2}$ *such that*

$$\sup_{j\in\{1,...,d_2\}}\left\|v_j - c_{vj} - \left(p^K\right)'\beta_{x,j}^K\right\|_{\infty,\mathcal{Z}} = O\left(K^{-\mu}\right),$$

as $N, T, K \to \infty$, *where* μ *is the constant in Assumption 1.*

(b) *There exists a positive constant* M^m *such that*

- $\sup_{t\in\{1,...,T\}}\sup_{i\in\{1,...,N\}}\mathbb{E}\left[m\left(z_{it}\right)^2\right] \leqslant M^m$,
- $\sup_{t\in\{1,...,T\}}\sup_{i\in\{1,...,N\}}\mathbb{E}\left[\left\|x_{it}\right\|^6\right] \leqslant M^m$,
- $\sup_{t\in\{1,...,T\}}\sup_{i\in\{1,...,N\}}\mathbb{E}\left[\left\|\left(x_{it} - \mathbb{E}\left[x_{it} \mid z_{it}\right] - \mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right] - \mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right]\right)\epsilon_{it}\right\|^5\right] \leqslant M^m$, *and*
- $\sup_{t\in\{1,...,T\}}\sup_{i\in\{1,...,N\}}\mathbb{E}\left[\left\|x_{it} - \mathbb{E}\left[x_{it} \mid z_{it}\right] - \mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right] - \mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right]\right\|^6 |\{z_{is}\}_{s=1}^T, \alpha\right] \leqslant M^m$ *for any* α *and* $\{z_{is}\}_{s=1}^T$, *almost surely.*

(c) *There exist a sequence of constants* Π^x *such that* $\sup_{k\in\{1,...,K\}}\left\|\beta_{x,j:k}^K\right\| \leqslant \Pi^x$ *and*

with Assumption 8.3. To avoid extra notation, I keep them as the same.

$$\frac{\sqrt{T}\Pi^x\sqrt{K}}{\sqrt{N}} \to 0, \text{ as } N,T,K \to \infty.$$

5. *(a) There exists a* $\delta' \in \left(0, \frac{1}{2}\right)$, *such that*

   i. $\frac{T}{N^{\delta'}} \to 0$, *as* $N,T,K \to \infty$.

   ii. $\frac{\xi_K^2\sqrt{K}}{N^{\frac{1}{2}-\delta'}\sqrt{T}} \to 0$, *as* $N,T,K \to \infty$.

   iii. $\frac{N^{\delta'}\xi_K}{K^{\mu}} \to 0$, *as* $N,T,K \to \infty$.

   iv. $\frac{\xi_K\sqrt{K}\Pi_K}{N^{\frac{1}{2}-\delta'}} \to 0$, *as* $N,T,K \to \infty$.

   (b) $\frac{T\sqrt{K}\xi^2\Pi_K}{\sqrt{N}} \to 0$, *as* $N,T,K \to \infty$.

   (c) $\frac{\sqrt{NT}\xi_K}{K^{\mu}} \to 0$, *as* $N,T,K \to \infty$, *where* μ *is the constant in Assumption 1.*

   (d) $\frac{K\xi_K}{T} \to 0$, *as* $N,T,K \to \infty$.

Assumption 9 provides sufficient conditions to derive the asymptotic distribution of $\hat{\theta}$ and uniform consistency of $\hat{\alpha}_{\hat{g}_i t}$. The proof uses the usual least squared formula but this requires the solution of $\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right)$ to be in the interior. Theorem 1 and Assumption 9.1 provide $\left(\hat{\theta}, \hat{\beta}^K\right)$ being in the interior with asymptotic probability one. However, until now, $\hat{\alpha}_{gt}$ is shown only to be mean-squared consistent over the sample path. This result is not enough to force it into the interior. However, verifying the solution as the interior is simple in practice.

Assumption 9.2 specifies weak dependency conditions. In the cross-section, it assumes α as the only source of cross-correlation for (x_{it}, z_{it}) [4]. In time-series, the weak

[4]Potentially, the cross-section can allow weak dependency after conditioning on α. One possible extension is to use Lee and Robinson (2016)'s setup to model cross-sectional dependency. But, for simplicity, this weak dependency is not considered here.

dependency is described by mixing conditions. For example, (x_{it}, z_{it}) have the said alpha mixing properties if they are functions of independent processes with these alpha mixing properties.[5]

Assumption 9.3 strengthens the rank. The moments are defined conditionally because cross-sectional independence happens only conditional on α. But, conditional on α, the regressors q_{it} are not stationary. Hence, the condition bases on an average of over T.

From Assumption 9.4.a, the same series basis can uniformly approximate the conditional expectation of x_{it} - Qi (2000) uses a similar setup. Also, it assumes the conditional expectation of x_{it} is a homogeneous function over the cross-section. This restriction still allows x_{it} to be heterogeneous in expectation from the heterogeneity of distribution of z_{it} over the cross-section.

From Assumption 9.5, N is assumed as larger than T to ignore the incidental parameter problem of α_{gt}. So 9.5.b provides the rate for it to happen. Again, Assumption 9.5.c assumes m is sufficiently smooth such that scaling by $\sqrt{NT}$ still leaves the approximation error to be asymptotically negligible. Newey (1997) and Qi (2000) provide their analogous versions of Assumption 9.5.c to derive the asymptotic distribution. Assumption 9.5.a rates ensure the estimate of $\hat{\alpha}_{\hat{g}t}$ is uniformly consistent, over T, under the non-parametric estimation of m. However, Assumption 9.5.a is only relevant for the two-step problem and can be ignored for $\hat{\theta}$'s asymptotic normality. Assumption 9.5.d rate ensures the θ's asymptotic covariance matrix is convergent to its population analog under the non-parametric estimation of m. However, it is not needed to derive the consistency rate provided in the next theorem - thus, the

[5]For reference, Andrews (1983) provides conditions to when a stationary autoregressive process is alpha mixing.

two-step estimation can ignore this rate.

Theorem 4. *(Asymptotic Normality and Uniform Convergence)*
Under Assumption 1-6 and 8-9,

1. $\sqrt{NT}\left(\hat{\theta}-\theta^0\right) \Rightarrow N\left(0,\Sigma_\theta\right)$ *where*

   a $\Sigma_\theta = \left(\sum_{g=1}^{G^0}\kappa_g\psi_g^{xz}\right)^{-1}\psi^{x\epsilon}\left(\sum_{g=1}^{G^0}\kappa_g\psi_g^{xz}\right)^{-1}$,

   b $\psi_g^{xz} = \lim_{N\to\infty}\left[\frac{\sum_{i:g_i^0=g}\mathbb{E}\left[\left(x_{i1}-\psi^{xx\epsilon}\left(z_{i1},\alpha\right)\right)\left(x_{i1}-\psi^{xx\epsilon}\left(z_{i1},\alpha\right)\right)'\right]}{N_g}\right]$,

   c $\psi^{x\epsilon} = \lim_{N\to\infty}\frac{\sum_{i=1}^{N}\left[\sum_{t=1}^{\infty}\mathbb{E}\left[\left(x_{i1}-\psi^{xx\epsilon}\left(z_{i1},\alpha\right)\right)\left(x_{it}-\psi^{xx\epsilon}\left(z_{it},\alpha\right)\right)'\epsilon_{i1}\epsilon_{it}\right]\right]}{N}$,

   d *and* $\psi^{xx\epsilon}\left(z_{it},\alpha\right) = \mathbb{E}\left[x_{it}\mid z_{it}\right]+\mathbb{E}\left[x_{it}\mid\alpha\right]-\mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right]$.

2. $\sup_{i\in\{1,\ldots,N\},t\in\{1,\ldots,T\}}\left|\hat{\alpha}_{\hat{g}_it}-\alpha^0_{g_i^0t}\right| = o_p(1)$,

3. *and* $\left\|\hat{m}-m\right\|_{\infty,\mathcal{Z}} = O_p\left(\xi_K K^{-\mu}\right)+O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right)$.

as $N,T,K\to\infty$.

Theorem 4 provides the asymptotic normality for $\hat{\theta}$'s inference and α's estimates as uniformly consistent. Strengthening Corollary 1, Theorem 4 provides consistency of $\hat{\alpha}$ for everyone at every period. For the nonparametric estimate, the terms $O_p\left(\xi_K K^{-\mu}\right)$ and $O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right)$ control the approximation and estimation errors, respectively. When moments are heterogeneous even within groups, the convergence rate has an extra term and the covariance matrix involves the group averaged $\mathbb{E}\left[x_{it}\mid\alpha\right]$ instead. The details are provided in the Appendix-Chapter 2.

The proof strategy of Theorem 4 relies on the asymptotic equivalence result im-

plied by Theorem 3. Furthermore, Theorem 3's asymptotic equivalence implies that the classification problem leads to no efficiency loss in the limit. For θ's estimator, Robinson (1988) considers the semiparametric efficiency bound as the variance of θ's non-linear least squared (NLLS) estimator θ when m is a known parametric function identified by a finite-dimensional parameter γ_m. Then Robinson (1988) shows the double-residual semiparametric regression obtains this efficiency bound when $\mathbb{E}\left[x_{it}|z_{it}\right] = \dfrac{\partial m\left(z_{it};\gamma_m\right)}{\partial \gamma_m}$, almost surely.

The same exercise can be done here under no serial correlation, conditional homoskedasticity, x_{it} as strictly exogenous. In the presence of serial correlation or conditional heteroskedasticity, the NLLS's inefficiency is well-known. Moreover, including lags of x_{it} can also improve the NLLS estimator's efficiency when x_{it} is just sequentially exogenous. However, under those three conditions, the NLLS estimator is sensible a benchmark because it achieves the Gauss-Markov condition for its linear coefficient estimator.

Differencing the model by its group-level means turns the model into a simple partially linear model. By the Frisch-Waugh-Lovell theorem, the linear coefficient estimator obtained from applying NLLS on the demeaned model is identical to the version of NLLS applied to the original model. Then adapting Robinson (1988)'s observation and assuming m and its expectation are diferentiable in γ_m, the asymptotic semiparametric efficiency bound is

$$\sigma_\epsilon^2(\mathfrak{X}_1 - \mathfrak{X}_2\left[\mathfrak{X}_3\right]^{-1}\mathfrak{X}_2')^{-1}, \tag{2.9}$$

where $\sigma_\epsilon^2 = \mathbb{E}\left[\epsilon_{it}^2\right], \mathfrak{X}_1 = \mathbb{E}\left[\lim_{N\to\infty}\dfrac{\sum_{i=1}^N \breve{x}_{i1}\breve{x}_{i1}'}{N}\right], \mathfrak{X}_2 = \lim_{N\to\infty}\dfrac{\sum_{i=1}^N \mathbb{E}\left[\breve{x}_{i1}\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)'\right]}{N},$

$\mathfrak{X}_3 = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \mathbb{E}\left[\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)'\right]}{N}$, $\breve{x}_{it} = x_{it} - \mathbb{E}\left[x_{it}|\alpha\right]$, and $\breve{m}\left(z_{it};\gamma_m\right) = m\left(z_{it};\gamma_m\right) - \mathbb{E}\left[m\left(z_{it};\gamma_m\right)|\alpha\right]$. Moreover,

$$\Sigma_\theta = \sigma_\epsilon^2 \left[\lim_{N\to\infty} \frac{\sum_{i=1}^{N} \mathbb{E}\left[\left(x_{i1} - \psi^{x x\epsilon}\left(z_{i1},\alpha\right)\right)\left(x_{i1} - \psi^{x x\epsilon}\left(z_{i1},\alpha\right)\right)'\right]}{N}\right]^{-1} \tag{2.10}$$

under those three conditions. Now it is apparent that Σ_θ obtains the semiparametric bound when $\mathbb{E}\left[x_{it}|z_{it}\right] = \frac{\partial m\left(z_{it};\gamma_m\right)}{\partial\gamma_m}$, as in Robinson (1988). The efficiency bound argument easily extends to the case of heterogeneous moments within-group, as described in the extended Assumption 9.

Now I propose an estimator for Σ_θ under all the previous assumptions. I construct the sample analogs $\widehat{\psi_g^{xz}} = \frac{\sum_{i:\hat{g}_i=g}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'}{\hat{N}_g T}$ and $\widehat{\psi^{x\epsilon}} = \frac{\sum_i^N\left[\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{x}_{it}\tilde{x}_{is}'\hat{\epsilon}_{it}\hat{\epsilon}_{is}\right]}{NT}$ - where $\hat{\epsilon}_{it}$ denotes the partially linear model's residuals. Recall, $p^K\left(z_{it}\right)$ denotes the basis. $\tilde{x}_{it,d}$ is the residual from the least-squared projection of $x_{it,d} - \frac{\sum_{i:\hat{g}_j=\hat{g}_i} x_{jt,d}}{\hat{N}_{\hat{g}_i}}$ onto $p^K\left(z_{it}\right) - \frac{\sum_{j:\hat{g}_j=\hat{g}_i} p^K\left(z_{jt}\right)}{\hat{N}_{\hat{g}_i}}$, with $\hat{N}_g = \sum_{i=1}^{N}\{\hat{g}_i = g\}$.

Corollary 2. *(Covariance Estimator)*
Under Assumption 1-6, and 8-9, $\hat{\Sigma}_\theta = \left(\sum_{g=1}^{G^0}\frac{\hat{N}_g}{N}\hat{\psi}_g^{xz}\right)^{-1}\hat{\psi}^{x\epsilon}\left(\sum_{g=1}^{G^0}\frac{\hat{N}_g}{N}\hat{\psi}_g^{xz}\right)^{-1}$ *is a consistent estimator of* Σ_θ.

The covariance formula is the version of Arellano (1987)'s within-group estimator after accounting for the non-parametric estimation of $\hat{m}$. In their appendix, Bonhomme and Manresa (2015) also considers the within-group estimator for the linear model. Within the large N and T framework, Hansen (2007) shows the within-group estimator as consistent for the linear model. I extend this consistency result into the partially linear case. In section 4, I assess the covariance estimator's performance in

constructing confidence intervals for a dynamic panel model with heteroskedasticity.

2.5. Monte Carlo

This section conducts a partially linear dynamic panel model experiment to assess the proposed estimators' performance for inference on θ. In particular, I construct confidence intervals for θ^0 and compare their actual coverage against their nominal values. Furthermore, I also report the classification errors to illustrate the classification error vanishes with a larger T.

In contrast to fixed effects, the dynamic panel model with the grouped fixed effect does not have an incidental parameter bias. Subsequently, there is no need to adjust confidence intervals with finite sample bias corrections. Hence, it is useful to experiment with a dynamic panel model to illustrate this advantage under a finite sample.

The dynamic panel model is

$$y_{it} = \theta y_{it-1} + \phi\left(z_{it}\right) + \alpha^0_{g_i^0 t} + \epsilon_{it} : \theta = 0.5, \tag{2.11}$$

where ϕ is a standard normal probability density function. This setup is a toy model of income growth, y_{it}, dependent of the level of inequality, z_{it}, and institutional effects, $\alpha^0_{g_i^0 t}$. By modeling z_{it}'s effect through the normal density, the model implies either the lack of or excessive inequality is not desirable for growth. Intuitively, lack of inequality may stifle incentives to produce, and excessive inequality can impede innovative entrant firms to access resources.

There are four groups, and each has a stationary process α_{gt} with a unique mean as either 0, 0.25, 0.5, or 1. z_{it} is the sum of $\alpha_{g_i^0 t}$ and another autoregressive process with

zero mean. Furthermore, the model has heteroskedasticity $\epsilon_{it} \sim N\left(0, \min\{1, y_{it-1}^2\}\right)$, independently. Conditional on α^0, all processes are independent over the cross-section. Moreover, all first-order autoregressive processes are generated by standard normal innovations and have 0.7 as its autocorrelation coefficient. The data generating processes are initialised at the stationary values.

Each group has the same number of memberships, and ϕ is approximated by a cross-validated polynomial of z_{it}. The other parameters are estimated as described previously.

N_g	$T = 5$	$T = 10$	$T = 15$	$T = 20$
40	90.1% (12.35%)	94.6% (1.81%)	96.8% (0.27%)	95.4% (0%)
100	85.8% (19.43%)	96.7% (4.55%)	96.4% (1.03%)	96.8% (0.27%)
200	80.5% (19.82%)	96.1% (4.86%)	96.6% (1.17%)	96.5% (0.31%)

Table 1: Coverage Probability for 95% Nominal Confidence Interval for $\hat{\theta}$

N_g	$T = 5$	$T = 10$	$T = 15$	$T = 20$
40	83.9% (17.35%)	90.2% (3.52%)	92.3% (0.77%)	91.3% (0.12%)
100	79.9% (19.43%)	91.8% (4.55%)	91.5% (1.03%)	91.7% (0.27%)
200	73.8% (19.82%)	91.2% (4.86%)	92.1% (1.17%)	92.2% (0.31%)

Table 2: Coverage Probability for 90% Nominal Confidence Interval for $\hat{\theta}$

The simulated results are tabulated from one thousand trials. The parenthesis reports the average classification errors[6] up to the second decimal, and the simulation shows the coverage is close to the nominal value as the number of periods increase. Furthermore, the classification error also drops with the number of periods, just as the asymptotic theory predicts.

The reader may notice two patterns from the tables. First, the classification error tends to be higher with a larger N_g. A larger sample is more likely to populate

[6]The estimated groups are only identified up to a permutation. To quantify the classification error, I match the estimated group with its members' modal true group g .

the empirical distribution's tail. Furthermore, the tail observations are likely to be misclassified. This phenomenon explains the little increase of classification error with larger N_g.

Second, the coverage probability tends to be larger than the nominal value with larger sample size. Under more numerous observations, z_{it} has more variations to reveal the finite polynomial's approximation error on ϕ.

N_g	$K=1$	$K=3$	$K=5$	$K=7$
200	84.8%	94.9%	92.7%	91%

Table 3: Coverage Probability for 90% Nominal Confidence Interval for $\hat{\theta}$ when $T=30$

N_g	$K=1$	$K=3$	$K=5$	$K=7$
200	92.3%	97.9%	96.2%	96%

Table 4: Coverage Probability for 95% Nominal Confidence Interval for $\hat{\theta}$ when $T=30$

It appears that the approximation error is causing the upward distortion of the coverage probability. The tables show coverage moves towards the nominal value by increasing K. Hence, the simulation is consistent with the theory's asymptotic prediction.

2.6. Conclusion

This chapter introduces a semiparametric panel model with time-varying grouped fixed effects for inference and to reveal latent economic forces available for economic analysis. Furthermore, the chapter proposes an estimator based on series approximation, for computational ease, and K-mean clustering. Subsequently, I provide

sufficient conditions for asymptotic inference and consistent estimation.

The chapter points out two advantages of using grouped fixed effects. First, it does not suffer from incidental parameter bias; therefore, the empirical researcher can avoiding doing finite sample bias correction. Second, the lack of incidental parameter problem allows the dominating consistency rates to vanish with N. Consequently, the group fixed effects asymptotic theory allows more freedom in choosing the number of basis terms for sample having comparably larger N than T - contrasting to existing results from using fixed effect or interactive fixed effect when the incidental parameter bias is an issue. The additional freedom translates to lower finite sample bias coming from the nonparametric approximation by series expansion.

Grouped fixed effects obtain its latent heterogeneity estimates by pooling over the cross-section. Hence, in short panels, grouped fixed effects offers more precise estimates of heterogeneity as to interactive fixed effects. This model offers a less noisier platform for researcher to learn if latent heterogeneity can explain observable economic characteristics. In chapter 3, I illustrate how grouped fixed effects explain the firm's ability to export and compete for market shares.

The next section considers some brief extensions of the model.

2.7. Extensions

High-dimensional z_{it}: A major nonparametric estimation challenge is the model's complexity to rise exponentially over the dimension of z_{it}, i.e., the curse of dimensionality. For a single variable, a third order univariate approximating polynomial consists of fourth parameters. While for three variables, the third order approximating polynomials has thirty parameters. Hence, a tremendous amount of data is required to estimate the series' parameters when z_{it}'s dimension is large.

One useful dimension reduction technique is to impose the index restriction, i.e., $m\left(z_{it}\right) = h\left(z_{it}'\upsilon\right)$, where function h is univariate and υ is an additional vector parameter. Subsequently, the proposed esimator for $m\left(z_{it}\right)$ is the series approximation by $p^K\left(z_{it}'\upsilon\right)'\beta^K$. Now the partially linear model estimator minimizes the new least-squared criterion:

$$\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}, \upsilon\right) \in \underset{\theta\in\Theta, \beta^K\in\mathcal{B}^K, \alpha\in\mathcal{A}^{G\times T}, \gamma\in\Gamma_G^N}{\arg\min} \hat{Q}(\theta, \beta^K, \alpha, \gamma, \upsilon), \tag{2.12}$$

where $\hat{Q}\left(\theta, \beta^K, \alpha, \gamma, \upsilon\right) = \dfrac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta^K - p^K\left(z_{it}'\hat{\upsilon}\right)'\beta^K - \alpha_{g_it}\right)^2$. The estimator of m is $\hat{m}(z_{it}) = p^K(z_{it}'\hat{\upsilon})'\hat{\beta}^K$.

Under this index restriction, an additional dimension to z_{it} adds only one new parameter. So it dramatically avoids the curse of dimensionality problem. The chapter's Algorithm 1 easily extends to optimize the new least squared criterion. The major difference is to do nonlinear optimization of the least-squared criterion function in both steps.

Multidimensional Heterogeneity: The partially linear model can have additional latent heterogeneous group structures for either θ or m. For example, there can be separate group structures for m and α. That is to say, the new partially linear semiparametric panel model is,

$$y_{it} = x_{it}'\theta^0 + m_i\left(z_{it}\right) + \alpha_{it}^0 + \epsilon_{it}, i, = 1, ..., N, t = 1, ..., T. \tag{2.13}$$

Each panel unit has a group membership h_i^0, not necessarily related to g_i^0, to describe its nonparametric function m_i. There are H different groups for m and the

relationship is

$$m_{it} = \begin{cases} m_{1t} & h_i^0 = 1 \\ \vdots & \vdots \\ m_{Ht} & h_i^0 = H. \end{cases} \tag{2.14}$$

For this new model, the partially linear model estimator minimizes the new least-squared criterion:

$$\left(\hat{\theta}, \{\hat{\beta}^{Kh}\}_{h=1}^{H}, \hat{\alpha}, \hat{\gamma_G}, \hat{\gamma_H}\right) \in \underset{\theta \in \Theta, \beta^{Kh} \in \mathcal{B}^K, \alpha \in \mathcal{A}^{G \times T}, \gamma_G \in \Gamma_G^N, \gamma_h \in \Gamma_H^N}{\arg\min} \hat{Q}(\theta, \{\hat{\beta}^{Kh}\}_{h=1}^{H}, \alpha, \gamma, \upsilon), \tag{2.15}$$

where $\hat{Q}\left(\theta, \{\hat{\beta}^{Kh}\}_{h=1}^{H}, \alpha, \gamma, \upsilon\right) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta^K - p^K\left(z_{it}\right)' \beta^{Kh_i} - \alpha_{g_i t}\right)^2$. The estimator of m is $\hat{m}_i(z_{it}) = p^K(z_{it})'\hat{\beta}^{K\hat{h}_i}$.

Chapter 4 provides a general algorithm to solve for multidimensional latent group structure. Its description outlines how to solve for this specific least-squared criterion. Furthermore, Chapter 3's extension section provides an economic motivation also to model m having a latent group structure.

Chapter 3

Production Function Estimation with Heterogeneous Dynamics in Productivity

3.1. Introduction

Modeling firms as production functions provide measurements to answer policy questions, such as *changes in industrial productivity* from regulations or trade liberalization and *documenting markup* to gauge the economy's concentration of market power. And to have robust measurements hinges on accounting for differences among firms. With microdata, a widely practiced rule is to designate homogeneous productivity transitions by industry classification digits. However, recent empirical evidence suggests there is substantial firm heterogeneity within a narrowly defined industry digits. Indeed, the North American Industry Classification System group the giant Amazon and any other online retailer into the narrowest industry digit.

There is a growing body of evidence to suggest modeling heterogeneity within narrowly defined industry digits is fruitful. For example, Raval (2020) documents a large variation of the firm's input-ratio within the US retailer industry. And Kasahara, Schrimpf, and Suzuki (2017) and Lee, Stoyanov, and Zubanov (2019) finds a paradoxical serial correlation of surprise productivity shocks when assuming productivity

transition is homogeneous within manufacturing industries of Chile, Denmark, and Japan. This chapter address this empirical paradox with its proposed estimator.

Olley and Pakes (1995) first introduced investments as a proxy variable to control for unobserved productivity in estimating production functions. Subsequently, Levinsohn and Petrin (2003) suggests replacing investments with intermediate material inputs because investments appeared lumpy in empirical data. Overall, the proxy variable method garnered considerable interest in the empirical literature with microdata - it has helped to generate interest in measuring the rise of US aggregate markup (see Autor et al. (2019) and De Loecker, Eeckhout, and Unger (2018)). More recently, Ackerberg, Caves, and Frazer (2015) provides a more modern treatment of this method.

There is also a growing interest in augmenting heterogeneity in the proxy variable model. One avenue is to use the firm's wage as instruments - see Doraszelski and Jaumandreu (2018). However, as Gandhi, Navarro, and Rivers (2017a) noted, to have wage as a reliable instrument is challenging for typical production function datasets. Recently, Lee, Stoyanov, and Zubanov (2019) proposed adding fixed effects, but information loss is typically quite significant by using within-transformation - see Griliches and Mairesse (1998). Another avenue is to make stronger assumptions about the firm's decision environment, e.g., perfect competition output market (see Griliches and Mairesse (1998)) or two flexible inputs with a competitive inputs market (see Demirer (2019)). However, some empirical questions may prefer to relax these assumptions, e.g., to allow non-perfectly competitive markets.

This chapter proposes an alternative model of heterogeneity by combining Ackerberg, Caves, and Frazer (2015)'s Leontief restriction with grouped fixed effects. The Leontief restriction helps to explain the intermediate material's monotonicity assumption

even when markets are not necessarily perfectly competitive. And using grouped fixed effects introduces an additional layer of heterogeneity without using within-transformation, or wage as an instrument, or first-order conditions based on assuming flexibility of the firm's input choices. However, the Leontief restriction requires a somewhat perfect complementary relationship between intermediate material and primary inputs. As noted by Ackerberg, Caves, and Frazer (2015), this assumption is plausible for manufacturing industries, which are of interest in many empirical applications.

Under this alternative model, this chapter proposes a data-driven rule to model heterogeneous productivity transitions for a production function estimator based on the proxy variable approach. Firms within a group share the same productivity transition, but the econometrician does not observe the group structure. For empirical application and Monte Carlo simulation, I use the commonly used Cobb-Douglas production technology. However, my presented analysis covers general Hicksian neutral production technology. The econometric solution is providing an estimator to estimate the production function parameters and the group structure simultaneously.

Recently, Kasahara, Schrimpf, and Suzuki (2017) proposes a finite-mixture model to estimate group structure in production function estimation. There are two significant differences between this chapter's approach and theirs. First, their estimator is a maximum likelihood-based with normality assumptions. Here, this chapter's estimator doesn't assume a specific distribution for productivity - which is the typical setup for most textbook production function estimators. Second, they assume a CES demand system and market structure; however, some applications would prefer to make minimal assumptions about demand. For example, De Loecker and Scott (2017) argues for measuring markup with a production function, mainly because conventional

production function estimators are agnostic about the nature of firms' competition - unlike the standard demand approach for markup estimates. This chapter closely follows Ackerberg, Caves, and Frazer (2015)'s structural value-added identification, which can be agnostic to competition assumption - provided measured outputs and inputs are measured correctly.

Here, I propose a two-step estimator where the first-step estimation applies the non-parametric regression case of the second chapter's model. The procedure is familiar to the typical two-step proxy variable estimation. Indeed, the chapter introduces a computationally convenient extension for empirical researchers accustomed to the widely used proxy variable approach. Furthermore, this chapter shows how the extension follows from the structural value-added identification (Ackerberg, Caves, and Frazer (2015)) and provides sufficient conditions for its estimator's consistency.

It is instructive to outline this chapter's estimation procedure briefly to facilitate further discussions. In the first step, the second chapter's nonparametric regression filters out the ex-post productivity shocks (or measurement error) and heterogeneous productivity transition. Then the second-step uses nonparametric regression's conditional mean to construct a set of non-trivial moment conditions and optimize them to estimate the production function's parameters. The moment conditions are non-trivial because they nest an additional first-order Markovian nonparametric regression, also dependent on production function parameters.

Under homogeneous productivity transition, Olley and Pakes (1995) provides conditions for the second-step estimator's consistency when a kernel estimator estimates the Markovian regression. Also noted by Olley and Pakes (1996), applying series approximation for the Markovian regression is computationally easier, but they don't provide sufficient conditions to do so. Indeed, the empirical literature has moved to

estimate the Markovian regression with series approximation.

Chen, Linton, and Keilegom (2003) studies a general framework that nests the econometric problem here, and they provide asymptotic theorems based on high-level conditions. This chapter offers low-level conditions for series to verify Chen, Linton, and Keilegom (2003)'s high-level conditions for consistency. The considered asymptotics is large N and T while allowing N as comparably larger than T to be coherent with the second chapter's analysis. Lastly, this chapter conducts all asymptotic analysis under the introduced heterogeneous productivity transitions.

For the Monte Carlo simulation, I consider a modified Ackerberg, Caves, and Frazer (2015)'s DGP1 model used in their simulation to assess the proxy variable performance. This model preserves the proxy variable estimator's consistency even when firms' wages are heterogeneous, and the econometrician doesn't observe the wage schedule - a typical environment in applications. My main modification is to build different productivity transitions into DGP1 - differing by a latent group structure. Simulation evidence shows my estimator can have desirable finite sample performance even under small T. In particular, the classification error of group structure diminishes over a lower variance of ex-post productivity shocks and a smaller correlation of heterogeneous productivity transition. Lastly, the simulation also shows how the information criterion overwhelming selects the true number of groups under DGP1.

This chapter concludes with an empirical application on the Chilean manufacturing firms to show two insights. The literature has extensively used this dataset to benchmark production estimation, hence applying the dataset here provides a useful reference point. First, I show including heterogeneous productivity transitions addresses empirical evidence of misspecification in the proxy variable model. Recently, Kasahara, Schrimpf, and Suzuki (2017) and Lee, Stoyanov, and Zubanov (2019) points

out unaccounted heterogeneity within industry digits leads to the paradoxical serial correlation of supposed surprise productivity shocks. The serial correlation shows over 70% reduction after incorporating heterogeneous productivity transition, with the number of latent groups chosen by an information criterion.

Second, the estimated groups capture salient features about the firm: its ability to export and to compete for market shares. The dataset covers the period of Chilean Miracle growth, happening after its economic liberalisation policies. My estimates show higher productive firms are more able to export and compete for larger market shares. The findings suggest the liberalisation policies are promoting more efficient Chilean firms. The exercise shows how latent groups help to document the market structure's effects on firms for economic policy analysis.

The rest of the chapter is organised as follows. In Section 3.2, I set up the production function model with its proposed estimator establish the asymptotic analysis in Section 3.3. I present a Monte Carlo simulation for finite sample performance in section 3.4 and empirical analysis in section 3.5. Then conclusion happens in Section 3.6, then follows by extensions in Section 3.7.

3.2. Model and Estimation

The objective is to estimate a parametric production function under the simultaneity problem - the firm bases its input choices on productivity, which is unobserved by the econometrician. Neglecting the simultaneity problem induces biases in the estimates - usually known as the "transmission bias". First, I generalise section two's example to allow smooth Hicksian production technology and ω_{it} as a first-order Markov process. The next section's conditional method of moments problem analyses this application under a general combination of moments.

At the high-level, my setup generalises the proxy variable approach on estimating the firm's production function. The generalisation introduces four additional features. The proxy variable approach assumes that different firms' productivity processes are first-order Markov and independent of each other. Furthermore, the Markov transition function is identical. I generalise by *introducing cross-correlation in firms' productivity* (1) and *relaxing the first-order Markov assumption* (2). It is plausible to assume the cross-correlation is relevant in application because spillover effects happen from technological advancements. Finally, I *weaken the scalar unobservable assumption* (3) and *allow firms' productivity transition dynamics to differ* (4).

After setting up the estimation procedure, I discuss my identifying assumptions. For readers only interested in the econometric setup, it is sufficient to read just the *Setup and Estimation* and the *Conditional Method of Moments* parts in this section.

Setup and Estimation

The Hicksian Neutral technology specification defines productivity to have the same effect on the marginal product of capital and labour. So the firm's production function is expressed as

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g_i^0 t} + \omega_{it}\right) F\left(K_{it}, L_{it}\right), \tag{3.1}$$

where Y_{it}, K_{it}, and L_{it} are output, capital, and labour, respectively. The production function can be log-linearised to form:

$$y_{it} = f\left(k_{it}, l_{it}; \tilde{\tau}\right) + \alpha^0_{g_i^0 t} + \omega_{it} + \epsilon_{it}, \tag{3.2}$$

where $y_{it} = \log\left(Y_{it}\right), k_{it} = \log\left(K_{it}\right)$, and $l_{it} = \log\left(L_{it}\right)$. The econometric objective is to estimate the parameter $\tilde{\tau}$ when the parametric form of f is known. For Cobb-Douglas case, $f\left(k_{it}, l_{it}\right) = \beta_k k_{it} + \beta_l l_{it}$ with $\tilde{\tau} = (\beta_k, \beta_l)$ and, for Constant Elasticity

of Substitution case, $f(k_{it}, l_{it}) = \beta_1 \log(\exp(k_{it}\beta_2) + \exp(l_{it}\beta_2))$ with $\tilde{\tau} = (\beta_1, \beta_2)$.

As before, I assume there is a variable v_{it} to proxy ω_{it} after accounting for the firm's input choice of (k_{it}, l_{it}). That is $\omega_{it} = h(k_{it}, l_{it}, v_{it})$, for some unknown function h. A popular choice of v_{it} is the firm's intermediate material input in the recent applied literature. In the next part, I briefly discuss when the h function is independent of α^0. Furthermore, I provide an example of the h function that is independent of α^0 in the *Structural Examples* part.

The h function provides the reduced form regression,

$$y_{it} = m(k_{it}, l_{it}, v_{it}) + \alpha^0_{g_i^0 t} + \epsilon_{it}, \tag{3.3}$$

where the nonparametric function $m(k_{it}, l_{it}, v_{it}) = f(k_{it}, l_{it}; \tilde{\tau}) + h(k_{it}, l_{it}, v_{it})$. In the absence of α^0, the reduced form regression is a standard first-step regression in the extensively used proxy variable's method to estimate the production function. As usual in the literature, h is treated as nonparametric and thus f is not separably identifiable from h. Then the reduced form regression is a semiparametric regression and can be estimated by the partially linear model's method in Chapter 2 2. This estimation forms the first-step and separates ω_{it} from $\alpha^0_{g_i^0 t} + \epsilon_{it}$. Hence, this estimation step can also be referred as the filtering step.

Let $v_{it} = \omega_{it} - \mathbb{E}[\omega_{it} \mid \omega_{it-1}]$. For now, I assume the following moment conditions are valid:

$$\mathbb{E}\left[\begin{pmatrix} k_{it-1} \\ l_{it-1} \\ 1 \end{pmatrix} v_{it}\right] = 0 \text{ and } \mathbb{E}[\omega_{it}] = 0. \tag{3.4}$$

The first set of moments assumes the firm to forecast ω_{it} with just the information of

ω_{it-1} when the firm chooses capital and labour inputs at the period $t-1$. For now, I assume $\tilde{\tau}$'s dimension is less than four for these moment conditions to sufficiently identify it. In the next part, assumptions are presented and more valid moments appear to identify $\tilde{\tau}$ when it has higher dimensions.

Next, I discuss how to construct sample moment analogues to estimate $\tilde{\tau}$. Based on m's identity, $\omega_{it} = m\left(k_{it}, l_{it}, \upsilon_{it}\right) - f\left(k_{it}, l_{it}; \tilde{\tau}^0\right)$[1]. Upon having the estimate $\hat{m}$, guessing $\tilde{\tau}^0 = \tilde{\tau}$ leads to a guess of

$$\hat{\eta}_{it}\left(\tau\right) = \hat{m}\left(k_{it}, l_{it}, \upsilon_{it}\right) - f\left(k_{it}, l_{it}; \tilde{\tau}\right) - \nu, \tag{3.5}$$

where $\tau = \left(\tilde{\tau}, \nu\right)$. The constant ν appears from m's intercept as not separately identified from $\alpha^0_{g_i^0 t}$ in the filtering stage.

To recover υ_{it}, I estimate $\mathbb{E}\left[\omega_{it} \mid \omega_{it-1}\right]$ by minimizing the least-squared prediction error of $\hat{\omega}_{it}\left(\tau\right)$ based off the series of basis functions $b^L\left(\hat{\omega}_{it-1}\left(\tau\right)\right)$. With the estimated firm's Markov prediction as $\hat{R}\left(\hat{\omega}_{it-1}\left(\tau\right)\right) = \hat{\omega}_{it-1}\left(\tau\right)' \hat{r}^L\left(\tau\right)$ then

$$\hat{\upsilon}_{it}\left(\tau\right) := \hat{\omega}_{it}\left(\tau\right) - \hat{R}\left(\hat{\omega}_{it-1}\left(\tau\right)\right). \tag{3.6}$$

For sections 3.5 and 3.6, I use the Cobb-Douglas production function and, hence, $\tilde{\tau}$ has dimension two. At there, the second-step's General Method of Moment (GMM) criterion is

$$\hat{\mathfrak{M}}_{NT}\left(\tau\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\sum_{t=2}^{T} l_{it-1}\hat{\upsilon}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=2}^{T} k_{it-1}\hat{\upsilon}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=1}^{T} \hat{\omega}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=2}^{T} \hat{\upsilon}_{it}\left(\tau\right)}{T}\right)^2\right].$$

Thus the estimator

$$\hat{\tau} \in \arg\min_{\tau} \hat{\mathfrak{M}}_{NT}\left(\tau\right).$$

[1] $\tilde{\tau}^0$ stands for the true value of $\tilde{\tau}$.

For identification purposes, the criterion does not pool the moments over the cross-section. When cross-sectional units have non-identical distributions, pooling the cross-sectional moments can cause the criterion to have non-unique minima.

The assumptions and the relationship with the proxy variable method

Here, I present my assumptions about the firm's behaviour on the proxy variable. Then I compare them to the standard assumptions in the literature, as presented by Ackerberg, Caves, and Frazer (2015).

Assumptions:

1. (Exclusion): v_{it} is neither capital nor labor.
2. (Scalar Unobservable): $v_{it} = \mathfrak{g}_t(k_{it}, l_{it}, \omega_{it})$.
3. (Strict Monotonicity): $\mathfrak{g}_t$ is strictly increasing in ω_{it}.
4. (Time invariance): $\mathfrak{g}_t(k_{it}, l_{it}, \omega_{it}) = \mathfrak{g}(k_{it}, l_{it}, \omega_{it})$.
5. (First-Order Markov): Let $\mathcal{I}_{it}$ be the firm's information set capturing all the firm's knowledge at the end of period t. $\mathbb{E}[\omega_{it} \mid \mathcal{I}_{it-1}] = \mathbb{E}[\omega_{it} \mid \omega_{it-1}]$. Furthermore, ω_{it} is a zero-mean process.
6. ("Surprise" shock): $\mathbb{E}[\epsilon_{it} \mid \mathcal{I}_{it}] = 0$.

It is instructive to first see how these assumptions sets up the estimation. The combination of assumptions 2, 3, and 4 imply $\mathfrak{g}$ as invertible with respect to ω_{it}, conditional on k_{it} and l_{it}. Then the unknown h is $\mathfrak{g}^{-1}(k_{it}, l_{it}, v_{it})$.

Assumption 5 verifies the provided moment conditions because k_{it-1} and l_{it-1} are in the firm's information set $\mathcal{I}_{it-1}$. Furthermore, Assumption 5 provides additional

moment conditions,

$$\mathbb{E}\left[\begin{pmatrix} k_{it-s} \\ l_{it-s} \\ 1 \end{pmatrix} v_{it}\right] = 0 \text{ and } \mathbb{E}\left[\begin{pmatrix} k_{it-s} \\ l_{it-s} \\ 1 \end{pmatrix} (v_{it} + \epsilon_{it})\right] = 0, \text{ for } s \geqslant 1, \tag{3.7}$$

because the further lags are also in $\mathcal{I}_{it-1}$. The second set of moments can be constructed by using the sample analogue,

$$\widehat{v_{it} + \epsilon_{it}}(\tau) = y_{it} - \hat{m}(k_{it}, l_{it}, v_{it}) - f(k_{it}, l_{it}; \tilde{\tau}) - \hat{\nu} - \hat{R}(\hat{\omega}_{it-1}(\tau)) - \hat{\alpha}_{\hat{g}_i t}. \tag{3.8}$$

These additional moments would help to identify $\tilde{\tau}$ when its dimensional is greater than four. The section's last part sets up the notation for the general method of moments problem.

In absence of α^0 (i.e. $\alpha^0_{g_i^0 t} = 0$), the first, second, third, and fifth assumptions are standard in the proxy variable literature. In the sixth assumption, I interpret ϵ_{it} as the unpredictable productivity shock, as first suggested by Olley and Pakes (1996). The fourth assumption assumes $\alpha^0_{g_i^0 t}$ to completely capture changes in the macroeconomic environment. This consequence is more restrictive than the general proxy variable framework - I call this the time-invariant proxy variable. Section 7[2] discusses on how to handle $\mathfrak{g}_t$ with finitely many structural changes over time. However, the time-invariant setup is the often adopted specification in practice.[3] The pertinent observation is my setup nests the time-invariant proxy variable model.

Logged capital investment (Olley and Pakes (1996)) and logged intermediate mate-

[2]This is located after the conclusion section.

[3]Allowing time-varying $\mathfrak{g}_t$ requires in splitting the observations to estimate multiple nonparametric functions.

rial (Levinsohn and Petrin (2003)) are two popular choices of v_{it} in the production function literature. When v_{it} is intermediate material, the function h is the firm's conditional demand of intermediate material. When v_{it} is the capital investment, the function h is the firm's investment demand.

Assumption 2 says firm's demand function of v_{it} is constant over $\alpha^0_{g_i^0 t}$, after conditioning on $(k_{it}, l_{it}, \omega_{it})$. As an example, this Assumption 2 holds for intermediate material when the firm's capital and labor input choices define its production capacity. Then the firm uses intermediate material to fill up its production capacity. ω_{it} can be understood as productivity that scales up the firm's capacity while $\alpha^0_{g_i^0 t}$ does not.

Say, for instance, the firm produces twenty defected units of goods for every two hundred in production. However, the firm can only sell its non-defected units. With better training and quality control, the firm can reduce its defect rate down to five percent; then, this change is a productivity increase. However, the amount of intermediate material used to produce each unit remains unchanged, and then $\alpha^0_{g_i^0 t}$ captures this productivity increase. More discussions about h as constant over α^0 are provided in the *Structural Examples* part.

Under the presence of $\alpha^0_{g_i^0 t}$, the fifth assumption is a bit nuanced. There is an implicit assumption of ω_{it} as mean independent of α^0 conditional on ω_{it-1}. Relaxing this assumption is straightforward, but it is kept for simplicity. All forms of cross-correlation is to be absorbed by $\alpha^0_{g_i^0 t}$ and this leaves ω_{it} as independent over i.

The standard proxy variable model precludes dynamic cross-correlation in firms' productivity because of the fifth assumption, and $\alpha^0_{g_i^0 t} = 0$. It is to imagine the firm observing (at least partially) its competitors' productivity. So other firms' productivity information should be in the set $\mathcal{I}_{it}$. The fifth assumption says their information

is not helpful to predict tomorrow's ω_{it+1} beyond knowing today's ω_{it}. For application, this means the firm's competitors can not independently innovate with positive spillover effects for the industry.

In the absence of $\alpha^0_{g_i^0 t}$, the Scalar Unobservable assumption with strict monotonicity predicts the firm to always increase its input level of v_{it} by a higher overall productivity level. This prediction is a reasonable assumption in the competitive market but can fail when the firm has market power. For example, if technological progress helps the firm to reduce its waste of intermediate material, then the firm with market power may want to cut its intermediate material purchases to raise profits. Then the strict monotonicity fails. With $\alpha^0_{g_i^0 t}$, the firm does not need to increases its purchases in a strict fashion with overall productivity.

Finally, $\alpha^0_{g_i^0 t}$ does not need to be a first-order Markov process. By including $\alpha^0_{g_i^0 t}$, the econometrician can be somewhat agnostic about the productivity process' order of persistence. Furthermore, the firms' productivity transition functions can now be different because $\alpha^0_{g_i^0 t}$ differs among firms. So my setup generalises the first order Markov setup used in the proxy variable approach.

Structural Examples

Here, I provide some worked out examples of $\mathfrak{g}$ as not dependent of $\alpha^0_{g_i^0 t}$.

Intermediate Material - v_{it} as logged intermediate material

The first example is in the setup of the structural value-added model, which is the frequently used example to justify the proxy variable assumptions - see Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2017b). In that setup, F is the firm's "valued-added" production function but the firm has a gross production

function described by the Leontief specification,

$$Y_{it} = \exp\left(\alpha^0_{g_i^0 t} + \epsilon_{it}\right) \min\{\mathcal{C}\left(M_{it}\right), \exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\}, \tag{3.9}$$

where M_{it} is the intermediate material and $\mathcal{C}$ is strictly increasing. But $\alpha^0_{g_i^0 t}$ is the same constant for every firm in the usual structural value-added model. The structural value-added model assumes the data-generating process is driven by the firm's interior solution of the Leontief model.

Under the usual structural value-added model, the firm's marginal product of intermediate material is predictably constant over time. My extension generalises the structural value-added model by allowing the firm to predict changes in the marginal product of intermediate material over time. Next, it becomes apparent that the structural value-added model illustrates the previously described capacity narrative.

The firm's interior solution has $M_{it} = \mathcal{C}^{-1}\left(\exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\right)$ because of $\mathcal{C}$'s strict monotonicity. Here, the $\mathfrak{g}$ function is $\log\left(\mathcal{C}^{-1}\left(\exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\right)\right)$ as not dependent of $\alpha^0_{g_i^0 t}$. Furthermore, the interior solution also implies

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g_i^0 t} + \omega_{it}\right) F\left(K_{it}, L_{it}\right)$$ [4]

[4]Under the interior solution, the semiparametric regression is

$$y_{it} = \log\left(\mathcal{C}\left(M_{it}\right)\right) + \alpha^0_{g_i^0 t} + \epsilon_{it}.$$

Which makes it just a semiparametric model of just intermediate material M_{it}. However, by generalising the gross production function to

$$Y_{it} = \exp\left(\alpha^0_{g_i^0 t} + \epsilon_{it}\right) \min\{\mathcal{C}\left(M_{it}, K_{it}, L_{it}\right), \exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\},$$

can provide the semiparametric regression as

$$y_{it} = \log\left(\mathcal{C}\left(M_{it}, K_{it}, L_{it}\right)\right) + \alpha^0_{g_i^0 t} + \epsilon_{it},$$

under the interior solution. Then the regression is a function of (M_{it}, K_{it}, L_{it}).

So the structural value-added model assumes the data generating process is based on firms applying the interior solution. When $\mathcal{C}$ is convex then it overcomes many concerns raised by Gandhi, Navarro, and Rivers (2017b) about the firm achieving the interior solution.

Investment - υ_{it} as logged investment

Economic models frequently assume that capital is subjected to some adjustment cost or delay with the installation. Hence, the firm's investment decision h is not sensitive to short-term productivity changes. Then $\alpha^0_{g_i^0 t}$ can stand as short-term productivity fluctuations. Furthermore, ω_{it} can stand for more persistent productivity changes.

In this environment, the firm's capital input is not correlated with $\alpha^0_{g_i^0 t}$. However, when the firm's labour input faces no dynamic constraints; labour is correlated with $\alpha^0_{g_i^0 t}$. For concreteness, I consider an example of Cobb-Douglas technology and a price-taking firm with its period t's capital investment as only effective at the start of period $t+1$. With the predetermined capital K_t, the firm chooses labour L_{it} to maximize its profit $\Pi_t(K_{it}) = p\mathbb{E}[\exp(\epsilon_{it})]\exp\left(\alpha_{g_i^0 t} + \omega_{it}\right)\left(K_{it}^{\beta_k} L_{it}^{\beta_k}\right) - rK_{it} - wL_{it}$, where p and (w, r) are output price and factor prices, respectively.

The standard optimization yields logged labour as $l_{it} = c_{Lit} + \dfrac{\alpha^0_{g_i^0 t}}{1-\beta_l} + \dfrac{\beta_k}{1-\beta_l} k_{it}$, where $c_{Lit} = \dfrac{1}{1-\beta_l}\log\left(\dfrac{p\beta_l}{w}\mathbb{E}[\exp(\epsilon_{it})]\exp(\omega_{it})\right)$. Hence, conditional on k_{it} and ω_{it}, l_{it} is still mean-dependent of $\alpha^0_{g_i^0 t}$. After the labour choice, the firm invests in capital, $\exp(\upsilon_{it})$, to maximize its discounted δ_1 future profit, $\sum_{T=t}^{\infty} \delta\mathbb{E}[\Pi_{t+1}(K_{it+1})|\mathcal{I}_t]$ subjected to the capital accumulation dynamic, $K_{is+1} = (1-\delta_2)K_{is} + \exp(\upsilon_{is})$, where $s \geqslant t$ and δ_2 is the depreciation rate. Suppose $\alpha^0_{g_i^0 t}$ is independently and identically distributed over time within the group. Then the future profit is constant over $\alpha^0_{g_i^0 t}$

and, in turn, the υ_{it} is constant of $\alpha^0_{g^0_i t}$. Thus the investment function $\mathfrak{g}$ does not depend on $\alpha^0_{g^0_i t}$. Finally, Olley and Pakes (1996) discusses how υ_{it} can be monotonic with respect to ω_{it} in the setup here.

Comparison Against the Alternatives

The fixed effects model is the first proposed solution to address this simultaneity problem. However, it requires the observed productivity to be time-invariant. Here, none of the firm's productivity components has to be time-invariant. Furthermore, the fixed effects estimator is known to produce unreasonably low capital coefficient estimates, as reviewed by Griliches and Mairesse (1998). The suspect is the fixed effects' within-transformation exacerbates attenuation bias from classical measurement error in the capital.

Griliches and Hausman (1986) shows attenuation bias increases as information is swept out of the regressors. The fixed effects estimator induces within-transformation, and the information loss is most severe when the regressors are highly serially correlated. As noted by Levinsohn and Petrin (2003), many firms make lumpy capital-investment decisions, and, as a consequence, capital is likely to be highly serially correlated. Fortunately, the grouped fixed effect estimator avoids within-transformation but applies between-transformation. Hence, the grouped fixed effects estimator is more resilient against attenuation bias, compared to fixed effects, when the between-firm variation is significantly larger than the within-firm variation. Section 6 re-visits this point and shows the between-firm variation is more pronounced in the Chilean data.

As an alternative to the proxy variable setup, the dynamic panel approach avoids the inversion setup, but it assumes the firm treats ω_{it} as an autoregressive process.

Furthermore, it estimates the autoregressive process with moment conditions. In summary, the dynamic panel method avoids the proxy variable assumptions for a simple autoregressive ω_{it} and using more moment conditions. More recently, Cheng, Schorfheide, and Shao (2019) shows how to estimate the dynamic panel approach with heterogeneous productivity means at the group level. In contrast to the fixed effects model, the group specification does not suffer the incidental parameter bias problem.

The other traditional avenues are to use either the firm's first-order condition behaviour or input prices as instruments. Imposing the firm's first-order condition either requires assuming perfect competition or the knowledge of each firm's output demand curve. Assuming perfect competition is not appropriate in applications where firms have market power, as in De Loecker, Eeckhout, and Unger (2018), De Loecker and Scott (2017), and De Loecker and Warzynski (2012). Recovering the firm's output demand curve requires additional consumer preference assumptions and the demand side's data set. For the input prices instrument approach, the firm's specific input prices must be available and provide valid exogenous variation. As both Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2017a) notes, having valid and reliable instrumental input prices is not typical in the data. In summary, these alternatives place a much higher demand for what is available in the data.

The proxy variable's niche is the combination of allowing a general Markov process, being a minimalist in both data requirement and making assumptions on the market structure, and utilising cross-sectional variation to control for ω_{it}. My extension introduces firms' correlated productivity while keeping many of the proxy variable's advantages.

Conditional Method of Moments

To cover the general production function problem, I set up a conditional method of moments problem. The interest is to estimate the parameter τ ($\in \mathcal{T}$) and its observable variables are generically denoted as w_{it} ($\in \mathfrak{W} \subset \mathbb{R}^{d_5}$) - potentially to include y_{it}, x_{it} or z_{it} and their lagged values.

Define $\underline{\omega}(z_{it}, \tau) := m(z_{it}) - f(z_{it}, \tilde{\tau}) - \nu$. Then the firm's Markov prediction can be expressed as,

$$R(\underline{\omega}(z_{it}, \tau), \tau) = \mathbb{E}\left[\underline{\omega}(z_{it+1}, \tau) \mid \underline{\omega}(z_{it}, \tau)\right]. \tag{3.10}$$

Since z_{it} is stationary from the partially linear theory's assumption, the function R is not a function of t. R is the first order autoregression of $\underline{\omega}(z_{it}, \tau)$ and the estimation procedure applies basis approximation to estimate R. Without loss of genearality, the R function can be treated as a function of (w_{it}, τ). Then, τ^0 solves the set of moment conditions,

$$\mathbb{E}\left[\mathfrak{m}\left(w_{it}, \theta^0, m(z_{it}), \alpha^0_{g_i^0 t}, \tau, R(w_{it}, \tau)\right)\right] = 0. \tag{3.11}$$

This setup covers moments built from higher ordered lagged inputs. The criterion function is denoted as,

$$\hat{\mathfrak{M}}_{NT}(\tau) = P'WP$$

where $P = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathfrak{m}\left(w_{it}, \hat{\theta}, \hat{m}(z_{it}), \hat{\alpha}_{\hat{g}_i t}, \tau, \hat{R}(w_{it}, \tau)\right)\right)$ and for some non-stochastic[5] positive definite weight matrix W. Then the second-step estimator

$$\hat{\tau} \in \arg\min_{\tau \in \mathcal{T}} \hat{\mathfrak{M}}_{NT}(\tau). \tag{3.12}$$

[5]Extending W to be stochastic is straightforward when W is asymptotically convergent in probability to a positive definite matrix. For simplicity, I omit this extension.

In the next section, I provide sufficient conditions for the consistency of $\hat{\tau}$. The strategy is to verify Chen, Linton, and Keilegom (2003)'s high-level assumptions for their Theorem 1 in my setup, where there are both time and cross-sectional dependence. Furthermore, I need to show $\hat{R}$ as uniformly consistent over w_{it} and τ. The problem is non-trivial because $\hat{R}$ is estimated by series where both its outcome and regressor depend on the parameter τ.

3.3. Asymptotic Theory

Sufficient conditions for $\hat{\tau}$

Here, I provide the sufficient conditions for consistency of the second-step estimator. The objective is to estimate τ in presence of the nuisance parameter $\mathfrak{h}(w_{it}, z_{it}, \tau) = (\theta, m(z_{it}), \alpha_{NT}(g), R(w_{it}, \tau))$, where $\alpha_{NT}(g) = \left\{\{\alpha_{g_{it}}\}_{t=1}^{T}\right\}_{i=1}^{N}$. In this notation,

$$\alpha_{NT}^{0}\left(g^{0}\right) = \left\{\left\{\alpha_{g_i^0 t}^{0}\right\}_{t=1}^{T}\right\}_{i=1}^{N}, \hat{\alpha}_{NT}(\hat{g}) = \left\{\{\hat{\alpha}_{\hat{g}_{it}}\}_{t=1}^{T}\right\}_{i=1}^{N},$$

$\hat{\mathfrak{h}} = \left(\hat{\theta}, \hat{m}(z_{it}), \hat{\alpha}_{NT}(\hat{g}), \hat{R}(w_{it}, \tau)\right)$, and $\mathfrak{h}^{0}(w_{it}, z_{it}, \tau) = \left(\theta^{0}, m^{0}(z_{it}), \alpha_{NT}^{0}\left(g^{0}\right), R^{0}(w_{it}, \tau)\right)$. Besides R, the nuisance parameters are inherited and estimated from the first step partially linear model. Here, I assume estimator of R as uniformly consistent but the next part provides the sufficient conditions for it to happen. Appendix - Chapter 3 contains the proof.

Assumption S 1. *(Identification)*

1. *For any $\delta > 0$, there exists an $\epsilon(\delta) > 0$ such that*

$$\inf_{\|\tau - \tau^{0}\| > \delta} \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^{0}(w_{it}, z_{it}, \tau)\right)\right]' \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^{0}(w_{it}, z_{it}, \tau)\right)\right] > \epsilon(\delta),$$

for any i and t.

In the model's setup, the true parameter τ^0 solves the conditional moments. Assumption S 1 ensures τ^0 does not suffer the weak identification issue with using the conditional moments.

Assumption S 2. *(Compactness)*

1. *$\mathcal{T}$ is compact and has a non-empty interior containing τ^0.*

2. *The supports $\mathfrak{W}$ and $\mathcal{Z}$ are compact.*

Assumption S 2.1 precludes the analysis in dealing with the boundary value problem. As in Chen, Linton, and Keilegom (2003), I require the sample analogs of the conditional moments to converge to its population criterion uniformly. The Assumption S 2.2 compactness assumption helps to ensure this convergence can happen in the N as comparably larger than the T paradigm. For the application, the compactness does not introduce new constraints. There is no additional random variable w, and the basis function is a polynomial. Generally speaking, polynomials can only achieve uniform approximation over compact sets.

Furthermore, the compact supports also help to turn the criterion into Lipschitz. The Lipschitz condition is a convenient assumption to achieve uniform convergence, as mentioned by Chen, Linton, and Keilegom (2003).

Assumption S 3. *(Moments and Dependency)*

1. $\left\{\left(w_{it}, z_{it}, \alpha^0_{g_i^0 t}\right)\right\}_{t=1}^{\infty}$ *has an alpha mixing coefficient $\rho_i^{w,z,\alpha}(t)$ satisfying*

$$\sup_{1 \leqslant i \leqslant T} \rho_i^{w,z,\alpha}(t) < C^{w,z,\alpha} \exp\left(-p_1 t\right),$$

for some constant $C^{w,z,\alpha}$ *and* $p_1 > 0$.

2. *For each* i, $\left(w_{it}, z_{it}, \alpha^0_{g_i^0 t}\right)$ *is a stationary process over* t.

Assumption S 3.1 is a weak dependency for the joint distribution of $\left(w_{it}, z_{it}, \alpha^0_{g_i^0 t}\right)$. The partially linear model does not necessarily have w_{it}. Hence, the previous weak dependency conditions alone do not necessarily imply Assumption S 3.1. z_{it} as a stationary process is already covered by Assumption 9, but is re-stated in Assumption S 3.2 for the ease of reference in the proof.

For the production function setup, these weak dependency conditions hold if intermediate material (or investment), labour, and capital are functions of independent state variables satisfying these weak dependency conditions. The simplest example is when the firm has no dynamic constraints on input choices and faces prices that are mutually independent processes and weakly time dependent.

However, it is natural to assume that capital faces dynamic constraints, then capital is also a state variable but it has a natural autoregressive transition. Then capital can be weakly dependent if the firm's investment function is sufficiently weakly time dependent. Furthermore, all state variables are no longer mutually independent because of capital as an additional state variable. So checking the weak time dependency for labour and intermediate material become more involved than before. It is useful to come up with simple sufficiency conditions for future research.

The estimation problem feeds the conditional moments with the first-step estimators, rather than the actual parameters. The estimation error is measured by the following metric,

$$d\left(\mathfrak{h}, \mathfrak{h}'\right) = \left|\theta - \theta'\right| + \sup_{z \in \mathcal{Z}} \left|m\left(z\right) - m'\left(z\right)\right| + \sup_{1 \leqslant i \leqslant N; 1 \leqslant t \leqslant T} \left|\alpha'_{g'_i t} - \alpha_{g_i t}\right| + \sup_{\tau \in \mathcal{T}} \sup_{w \in \mathcal{W}} \left|R\left(w, \tau\right) - R'\left(w, \tau\right)\right|.$$

Assumption S 4. *(Regularity)*

1. R^0 *is continuously differentiable.*

2. m^0 *is continuous.*

3. $\mathfrak{m}$ *is continuously differentiable over* $\mathbb{R}^{3+d_2+d_3+d_4}$.

These regularity conditions and the previous compact support assumptions complete $\mathfrak{m}$ as Lipschitz. For the production function case, differentiability is easy to verify for m^0 and $\mathfrak{m}$. For example, m^0 is differentiable when the parametric production function is differentiable, and the firm's conditional demand of the proxy variable has a nowhere vanishing derivative.

Assumption S 5. *(Rate)*

1. $\dfrac{\log(N)}{T} \to 0$, *as* $N, T \to \infty$.
2. $d\left(\hat{\mathfrak{h}}, \mathfrak{h}^0\right) = o_p(1)$.

Assumption S 2.2 is immediate from the results in Theorem 4 and the presumption of having a consistent estimator for R^0. Assumption S 2.1 is compatible with the larger N than T setup as the log function is slowly varying.

Theorem S 1. *Under Assumption 2, S 1, S 2, S 3, S 4, and S 5,*

$$\hat{\tau} \xrightarrow{P} \tau^0,$$

as $N, T \to \infty$.

The case of a stochastic matrix W is not formally covered here. However, Theorem S 1's argument can be adapted to hold when the stochastic W satisfies the high-level

conditions specified in Chen, Linton, and Keilegom (2003)'s Corollary 1.

The paper currently does not provide the theory to make inference on $\hat{\tau}$. Verifying the sufficient conditions leading up to Chen, Linton, and Keilegom (2003)'s Theorem 2 would provide a central limit theorem result. One important condition is to have $\hat{\mathfrak{h}}$ to converge at the $(NT)^{\frac{1}{4}}$ rate. The partially linear theory shows that this can happen for $\hat{m}$ and $\hat{\theta}$. Furthermore, the next subsection provides sufficient conditions for $\hat{R}$ to do so. However, $\hat{\alpha}_{\hat{g}_i t}$'s rate is unknown and to proceed forward may require dropping moment conditions using $\hat{\alpha}_{\hat{g}_i t}$. Furthermore, Chen, Linton, and Keilegom (2003) also requires a Donsker condition on the criterion function, and they only provide a reference to verify this condition for cross-sectional data with independence. However, section 4 shows that the bootstrap confidence interval provides the correct coverage in simulation when T is large. It appears the normality approximation and bootstrap standard errors can be used for inference, even under the cross-sectional dependence from α^0 and the time-series dependence for each unit.

3.4. Monte Carlo

Here, I assess the finite sample performance of my production function estimator when the intermediate material acts as the proxy variable. Also, the Monte Carlo verifies the asymptotic of my information criterion and classification consistency results when a polynomial non-parametrically estimates m.

My data-generating process is an extension of Ackerberg, Caves, and Frazer (2015)'s DGP1 used in their Monte Carlo. Their setup provides a simple solution to the firm's dynamic profit maximization problem and, then, simulates the data from firms' policy functions.

The structural value-added production function is Cobb-Douglas and the gross pro-

duction function is

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g^0_i t}\right) \min\left\{\mathcal{C}\left(M_{it}\right), \exp\left(\omega_{it}\right) K_{it}^{\beta_k} L_{it}^{\beta_l}\right\}, \tag{3.13}$$

where $\mathcal{C}\left(M_{it}\right) = M_{it} + M_{it}^2$. The objective is to estimate the output elasticity (β_k, β_l). $\mathcal{C}\left(M_{it}\right)$'s monotonicity and convexity ensures the firm's interior solution. Furthermore, I can easily solve the optimal choice of M_{it} from $\mathfrak{F}\left(M_{it}\right)$'s second-order polynomial form. Ackerberg, Caves, and Frazer (2015) used $\mathfrak{F}\left(M_{it}\right)$ as a linear function of M_{it} but, after log-linearization, the filtering step's semiparametric form is exactly linear in logged M_{it}. From a non-parametric perspective, it is uninteresting to approximate logged M_{it} with a polynomial of itself. However, the group productivity extension works perfectly fine with the linear specification.

Here, I outline the productivity process, and Appendix D provides the full firm's decision problem, solves the policy functions, and other parametric details. There are three true groups, i.e., $G^0 = 3$. All three processes ϵ_{it}, $\alpha^0_{g^0_i t}$, and η_{it} are mutually independent. Both ϵ_{it} and η_{it} are zero-mean independent processes over i but only ϵ is independent over t. η_{it} is a stationary first-order autoregressive process. The simulation generates firms' input choices based on the solved policy functions.

After applying the log transformation to the firm's interior solution of intermediate material,

$$y_{it} = \log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right) + \alpha^0_{g^0_i t} + \epsilon_{it}, \tag{3.14}$$

where $\upsilon_{it} = \log\left(M_{it}\right)$. So I use a cross-validated polynomial of υ_{it} to non-parametrically estimate $\log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right)$. For simplicity, I estimate R parametrically in the second stage.

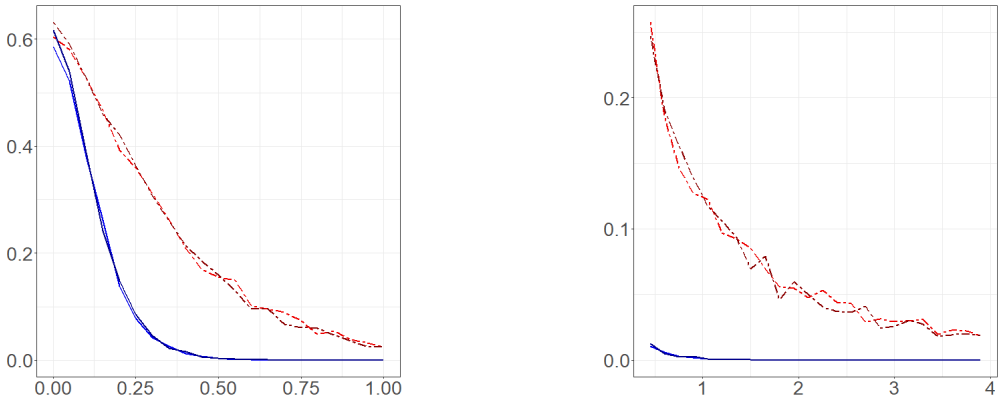
In the absence of $\log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right)$, the classification problem is on $\alpha^0_{g^0_i t} + \epsilon_{it}$. So intuitively, the classification of g^0_i is an easier problem when $\alpha^0_{g^0_i t} + \epsilon_{it}$ is more similar within the group than between groups. In my Monte Carlo setup, the classification precision is roughly positively associated with the ratio

$$\min_{g,g':g\neq g'} \frac{2\left(\sigma^2_{\alpha_g} - \sigma_{\alpha_g,\alpha_{g'}}\right) + \left(\mu_g - \mu_{g'}\right)^2}{\sigma^2_\epsilon}, \tag{3.15}$$

where $\mu_g, \sigma^2_{\alpha_g}$ and $\sigma_{\alpha_g,\alpha_{g'}}$, σ^2_ϵ are α_{gt}'s mean, variance, covariance with $\alpha_{g'}$, and ϵ_{it}'s variance, respectively. Appendix D provides a heuristic argument to why the ratio is informative in the partially linear semiparametric model.

The ratio suggests classification error decreases when different groups' α_{gt} become more dissimilar in mean or correlation. For my Monte Carlo Design 1, I model $\alpha^0_{gt} = w\alpha^*_{gt} + (1-w)\alpha^*_t$ - a convex combination of two mutually independent first-autoregressive processes, α^*_{gt} (group specific with $\mu_g \in \{-0.33, 0, 0.33\}$) and α^*_t (zero mean common trend). Classification error should fall as the grouped gross productivity processes become less correlated or more different in their means, i.e. when $w \to 1$.

The ratio also suggests the classification error decreases when $\frac{\sigma^2_{\alpha_g}}{\sigma^2_\epsilon}$ increases. That is to say, the classification precision improves when the firm's surprise productivity change ϵ_{it} comparably lowers in uncertainty. More predictable production environment should yield better classification estimates. For my Monte Carlo Design 2, I set $w = 1$ and vary $\frac{\sigma^2_{\alpha_g}}{\sigma^2_\epsilon}$ by increasing $\sigma^2_{\alpha_g}$. The Figure 1 estimates the average classification error at different parametric values, based on four hundred trials. And the simulated curves verify the previous predictions. Furthermore, the $T = 20$ curves are strictly lower than the $T = 5$ curves. Hence, they verify the asymptotic classifica-



Design 1 Design 2

Figure 1: Y-axis: Average Classification Error | X-axis: w and $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon}$ for Design 1 and 2, respectively.

Solid/Blue line: $T = 20$. Dashed/Red line: $T = 5$. Blue/Red: $N_g = 100$ (Number of observations for each group). Dark Blue/Dark Red: $N_g = 300$.

tion consistency result in the semiparametric model. Appendix D plots the difference between the mean-squared error of the elasticity estimates based on the true groups vs. the estimated groups. Their difference converge towards zero as the classification error vanishes. For Design 1, their difference also vanishes as $w \to 0$. Closer to $w = 0$, while classification identification is weaker, the elasticity estimates' bias is also smaller.

In the absence of classification error, production function literature has numerous studies of the two-step estimator's performance in Monte Carlo simulations. For brevity, I do not provide additional analysis of output elasticity estimates' mean-squared error, but next bootstrap results partially capture the estimator's performance in mean-squared error. For the inference of output elasticity estimates, I assess the bootstrap confidence interval's coverage. I do block bootstrap with each unit's time-series constituting a block. Each bootstrap sample constructs new output elasticity estimates at the second-step conditioning on the original sample's first-step

estimates. Then I construct the bootstrap confidence interval from the normal critical values and the bootstrap empirical distribution's[6] standard errors.

T	N_g	β_k^0	β_l^0
5	100	81.5%	64%
5	300	73.5%	48%
20	100	96.5%	93.5%
20	300	95%	96%

Table 5: Coverage for the 95% Bootstrap Confidence Interval.

The table reports the coverage probability over four hundred trials and under Design 1 with $w = 0.5$. The coverage converges to the nominal value as the number of periods increases. From the classification perspective, this outcome is not surprising as classification error is near zero at $T = 20$. However, the bootstrap confidence interval is able to provide the near correct coverage despite the data's serial correlation and cross-sectional dependence. As already mentioned, the bootstrap theory to account for both serial correlation and cross-sectional dependence is not provided here. But the simulation provides an applied justification to use bootstrap standard errors for section 5.

For the information criterion, I set the penalty as $\frac{\lambda}{T^{\frac{1}{5}}}\hat{Q}_{G_{\max}}$[7]. Then λ is chosen by the following data-driven approach:

$$\hat{G} \in \underset{G \in \{1,...,G_{\max}\}}{\arg\min} \; IC_{\lambda^*}(G), \tag{3.16}$$

where $\lambda^* \in \underset{\lambda \in \mathcal{K}}{\arg\min} \left[\underset{G \in \{1,...,G_{\max}\}}{\min} IC_\lambda(G) \right]$ and $\mathcal{K} := \{0.18, 0.2, ..., 1.8, 2\}$. The asymptotic theory covers criterion's selection consistency over every $\lambda \in \mathcal{K}$ because $\mathcal{K}$ is

[6] Five hundred bootstrap samples construct the empirical distribution.

[7] $\hat{Q}_{G_{\max}}$ is the least-squared criterion evaluated at the parameters estimated from the $G_{\max}$ specification. Having the penalty scaled by $\hat{Q}_{G_{\max}}$ ensures the selection is invariant to the data's scale.

a finite and fixed set. So the asymptotic result easily extends to the information criterion using the data-driven choice of λ.

At $w = 0.5$ for Design 1 or $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon} = 1$ for Design 2, the simulated classification error is around 17% when $T = 5$. With each group having 100 members, I assess the above information criterion's performance in the simulation at $w = 0.5$ for Design 1 and at $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon} = 1$ for Design 2. Here, $G^0 = 3$ and the $G_{\max} = 6$. All group specifications use the same polynomial order, and the order is chosen by cross-validation based on the over-specification, $G = 6$. How to optimally and jointly determine the polynomial order and the true group G^0 is an avenue for future research.

Design	T	$\hat{G} = 1$	$\hat{G} = 2$	$\hat{G} = 3$	$\hat{G} = 4$	$\hat{G} = 5$	$\hat{G} = 6$
1	$T = 5$	0%	2.9%	84.7%	12.4%	0%	0%
1	$T = 20$	0%	0%	99.5%	0.5%	0%	0%
2	$T = 5$	0%	0.5%	86.5%	13%	0%	0%
2	$T = 20$	0%	0.1%	99.4%	0.5%	0%	0%

Table 6: Simulated frequency of $\hat{G}$'s realisation based on four hundred simulations at each specification.

The table shows the information criterion performs well in finite sample even under classification error. Furthermore, the table verifies the asymptotic consistency result of the information criterion. The error of both under-selection and over-selection decreases as T increases. As this information criterion performs well here, I use it for my empirical application.

3.5. Empirical Analysis

In this section, I illustrate the empirical performance of my production function estimator. The data set consists of Chilean manufacturing plants from 1987 to 1996 and

is sourced from the census of Chilean manufacturing plants. It covers all firms with more than ten employees. My construction of capital, labor, and intermediate material follows Gandhi, Navarro, and Rivers (2017a, 2017b). For studying production function estimation, this data set series has also been used by Levinsohn and Petrin (2003) and Lee, Stoyanov, and Zubanov (2019).

The data set is available from 1979 to 1996. However, Levinsohn and Petrin (2003) raises potential structural break concerns for the earlier years. I use the data from 1987 to avoid addressing those structural breaks. The four sectors Food Product (331), Wood Products (331), Textile (321), and Fabricated Metal Products (381) are within the data set's top five largest sectors and included in all the mentioned previous studies. I restrict my analysis to these four sectors.

Chile experienced significant economic growth from 1987 to 1996, and the years fall into the well-known Miracle of Chile period. The growth spurt occurred after significant economic reforms were implemented and can be interpreted as the economy's convergence to a new steady state. Here, I use my production function estimator to measure firms' productivity changes and distribution for the four sectors. In contrast to previous studies, I allow cross-correlation in firms' productivity and firms to have heterogeneous transition dynamics in productivity.

Within a sector, I assume all firms have the same output elasticity (β_k, β_l) and follow the Cobb-Douglas (structural value-added) production function,

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \alpha^0_{g^0_0 t} + \epsilon_{it}, \tag{3.17}$$

with heterogeneous firm productivity. Using my production function estimator, I estimate the output elasticity for every four sectors. Here, the proxy variable is the

firm's intermediate material choice. I use a second order polynomial of $(k_{it}, l_{it}, \upsilon_{it})$ for the filtering step and a third order polynomial of $\hat{\eta}_{it-1}$ to approximate its first-order Markov process.[8] Using the second order at the filtering step is common in the literature because it is parsimonious and has the translog production function interpretation. As shown in section 4, I use the information criterion to select the number of groups for each sector - the set of alternatives includes up to ten groups, and the polynomial is the second order. The main estimates use four groups for Food, and five groups for Metal and Textile, and six groups for Wood.

Selection:

Both Olley and Pakes (1996) and Griliches and Mairesse (1998) argue for using the unbalanced panel to mitigate the selection issue from the firm's entry and exit decisions. Beyond using the unbalanced panel[9], I do not address the selection issue. It is possible to include a near-verbatim Olley-Pakes style selection correction at the second GMM step, but that is beyond the paper's scope.

Sector	N	Median T_i	Mean T_i
Wood	236	4	5.26
Textile	320	6	6.41
Food	1140	8	6.79
Metal	436	5	5.90

Table 7: *T_i is the ith firm's number of periods.*

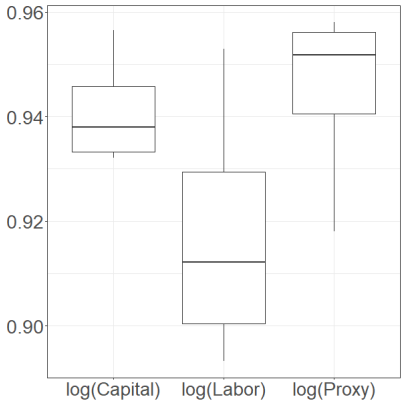
All sectors have a sizable N dimension comparably to their T dimension. In sec-

[8]For robustness, all local optimization steps are done with over five hundred randomly selected initialization points.

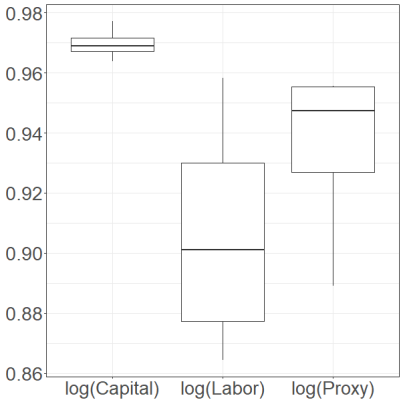
[9]It is not apparent on which T to substitute into the information criterion in this unbalanced panel setting. For simplicity, I use the firm's median number of periods.

tion 4, precise classification can be achieved even when T is small but N is large. More specifically, this happens when the firm's productivity uncertainty is low, i.e. ϵ_{it} is less variable than α_{gt}. Or, when the different mean-level of α_{gt} are well-separated.

Measurement Error: Within-Variation vs Between-Variation



R^2 of the logged input's AR(1) model.



$$\frac{\text{Logged Input's Variation Between Firms}}{\text{Logged Input's Total Variation}}$$

Figure 2: Sources of Inputs' Variation.

The Figure 2 shows that the inputs are highly serially correlated, and the firms' between-variation of inputs dominates the firms' within-variation of inputs. As discussed in section 3, Griliches and Hausman (1986)'s intuition suggests the attenuation bias is less severe by applying between-transformation as opposed to within-transformation in the scenario here. On the measurement error issue, I note that the grouped fixed effects estimator should be more resilient to the attenuation bias's effect, as compared to the fixed effects estimator.

Evidence of Heterogeneous Productivity Groups

I find the model specification's fit is improved by including heterogeneous productivity groups. The evidence lies with the estimates of ϵ_{it}.

The productivity ϵ_{it} is unaccounted by the firm's input decisions because it is unpredictable during the firm's decision making. From the filtering step, the estimate $\hat{\epsilon}_{it}$ is invariant for all smooth Hicksian neutral technology choice - only the second GMM step imposes the production function's parametric form. Furthermore, the filtering step's estimates are robust to the standard selection concern - which arises in the second GMM step. Hence, the estimate $\hat{\epsilon}_{it}$ is reasonably robust and should be serially uncorrelated under the correct model specification.

The time-invariant Proxy Variable model is nested under the single group specification. The single group is misspecified, as shown in Figure 3 from the $\hat{\epsilon}_{it}$'s significant autocorrelation. Moreover, the autocorrelation faces an over 70% reduction by increasing the number of groups to four or five. The Figure 3 is consistent with the presence of predictable grouped productivity shock in the firm's decision environment. For completeness, I also estimate the time-varying Proxy Variable model with the same second-order polynomial. Even in there, $\hat{\epsilon}_{it}$'s AR(1) model has a R^2 of $0.625, 0.748, 0.631$, and 0.707 for Wood, Food, Metal, and Textile, respectively.[10]

[10]On the over-fitting case, the time-varying Proxy variable model is more parameterized than my heterogeneous groups specification. My specification only adds one intercept per group for an additional year. However, the time-varying proxy variable model adds a new complete set of polynomial coefficients. My model is a parsimonious way to include heterogeneous productivity over time and different firms.

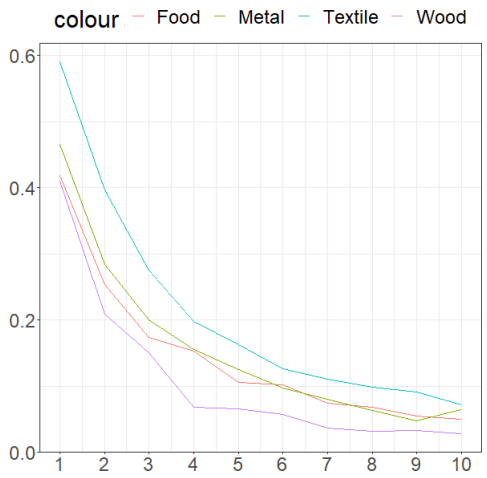


Figure 3: The R^2 of $\hat{\epsilon}_{it}$ AR(1) Model over G groups. Selected G is 4 for Food, 5 for Metal and Textile, and 6 for Wood.

The serial correlation of the proxy variable model's $\hat{\epsilon}_{it}$ has also been documented by Kasahara, Schrimpf, and Suzuki (2017), for the Japanese Machine Industry, and by Lee, Stoyanov, and Zubanov (2019), for Danish manufacturing firms.

As an alternative interpretation, Ackerberg, Caves, and Frazer (2015) models ϵ_{it} as serially correlated measurement errors on output. These measurement errors are innocuous only if they do not correlate with the measurement of inputs. However, I find the group with higher productivity tends to choose more capital input.[11] Capturing only innocuous measurement errors is not what drives down ϵ_{it}'s serial correlation in the graph. Nevertheless, the measurement error can be the cause behind the residual serial correlation after modeling the groups.

Groups' Composition:

[11] Stacked barplots of groups' mean level inputs are available in Appendix E.

Sector	G1	G2	G3	G4	G5	G6
Food	20.282	47.914	25.914	5.891	N/A	N/A
Metal	1.946	18.288	27.743	36.459	15.564	N/A
Textile	10.434	26.621	35.592	23.452	3.901	N/A
Wood	13.699	3.143	29.976	33.441	11.604	8.139

Table 8: Percentage of the sector's firm in each group. Groups are ordered in increasing mean level of $\hat{\alpha}_{gt}$.

Sector	G1	G2	G3	G4	G5	G6
Food	2.179	13.619	56.136	28.382	N/A	N/A
Metal	0.113	4.624	13.406	37.491	44.366	N/A
Textile	1.639	12.844	52.104	30.993	2.421	N/A
Wood	12.15	0.92	20.688	43.03	11.841	11.376

Table 9: Group's Market Share within Industry. Groups are ordered in increasing mean level of $\hat{\alpha}_{gt}$.

The table 8 shows that each estimated group is generally well populated. So the α_{gt} estimates use ample observations generally across the different groups. Appendix-E has stacked bar plots showing the differences in mean level input choices among the groups. For all sectors, the group's mean level of $\hat{\alpha}_{gt}$ increases with the group's mean level of capital. A costly capital adjustment model can explain this association. To avoid frequently adjusting capital, the firm front-loads its investment needs, and the level of front-loading increases with higher grouped productivity. With more flexible inputs, the firm weighs more on other short term aspects, from the demand side, in its input choices. This aspect explains why the mean level grouped productivity does not have a strict positive relationship with the mean level of intermediate material

and labour in the Textile, Wood, and Metal sectors. However, all three inputs hold a fairly positive association with grouped productivity over all sectors.

The table 9 shows that the market shares concentrate within a few groups more than what table 8's population count suggests. The low productivity groups have a disproportionately small market share relative to their firm population. Thus these sectors' aggregate output growths are more sensitive to the productivity changes in the highly productive groups. It may be interesting to match the groups with other observable characteristics to better understand the engine behind the Chilean economic growth in future research.

Output Elasticity Estimates and the Transmission Bias

Sector	$OLS: \hat{\beta}_k$[12]	$OLS: \hat{\beta}_l$	$G = 1$:$\hat{\beta}_k$	$G = 1$:$\hat{\beta}_l$	$G > 1$:$\hat{\beta}_k$	$G > 1$:$\hat{\beta}_l$
Food	0.341 (0.016)	0.815 (0.032)	0.33 (0.025)	0.539 (0.045)	0.3 (0.024)	0.517 (0.047)
Metal	0.219 (0.028)	0.917 (0.044)	0.225 (0.041)	0.667 (0.066)	0.151 (0.041)	0.657 (0.068)
Textile	0.233 (0.028)	0.78 (0.044)	0.192 (0.04)	0.72 (0.062)	0.176 (0.042)	0.7 (0.06)
Wood	0.195 (0.036)	0.975 (0.063)	0.156 (0.053)	0.895 (0.095)	0.181 (0.057)	0.829 (0.091)

Table 10: Output Elasticity Estimates - last two columns report the heterogeneous specifications. For the heterogeneous specification: $G = 4$ for Food, $G = 5$ for Metal and Textile, and $G = 6$ for Wood.

Heuristically, the transmission bias is positive for the elasticity estimate of the more flexible input. The firm prefers to adjust for more flexible input when productivity increases. Typically, labour is assumed to be a more flexible input than capital. In line with this theory, the table shows the OLS $\hat{\beta}_l$ is the largest. Then the estimate of β_l further decreases from the single productivity group specification to the heterogeneous

productivity group specification.

The heterogeneous groups' elasticity estimates have their return-to-scale hovering between 0.808 and 1.01. They are reasonably close to constant return-to-scale. The reported capital coefficient estimates are statistically significant from zero.[13] These estimates verify the conjecture of grouped fixed effects being more resilient to attenuation bias as compared to fixed effects.[14]

As already mentioned, the information criterion selects the number of groups here. In Appendix E, I plot output elasticity estimates for different G specifications. The output elasticity estimates are quite sensitive over different group specifications. Finding alternative methods to select the number of groups for the production function is an avenue for future research.

Productivity Heterogeneity and Productivity Growth

Here, I assess the difference in productivity measurement from accounting for heterogeneous productivity groups. The first part captures the difference in productivity growth's effect on output. Then the difference in productivity distribution's dispersion is examined.

[13]The statistical significance is at the 5% level if the bootstrap confidence intervals are valid. The paper has only studied the confidence intervals with 4's simulation.

[14]Lee, Stoyanov, and Zubanov (2019) estimated the output elasticity for the Chilean Textile sector. However, they treated α as a firm fixed effect. Their Textile sector's capital-output elasticity estimate is at a single-digit percentage point and statistically insignificant from zero at the 10% level. The usual suspect is attenuation bias from the capital's measurement error.

	$G = 1$	$G > 1$
Food	2.078	2.277
Metal	4.98	5.89
Textile	1.5	1.57
Wood	0.81	0.4

Table 11: Average Output Growth Due to Productivity - Controlling for Inputs Level

Using Olley and Pakes (1996) 's formula, I decompose the annual output growth due to productivity growth after controlling for inputs level. After averaging them over the years, I report them in table 11 for each sector. For the Metal sector, heterogeneous group specification accounts for at least 18 % more growth from productivity - 9.6 % for the Food sector. Interestingly, the Wood sector reports a lower rate under heterogeneity. Appendix E shows that the lowest productivity group experienced a sizeable productivity contraction for some time. Homogeneous specification hides this fact, and it may explain the difference.

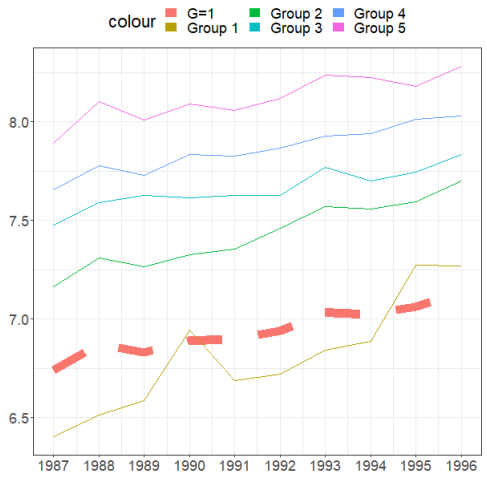


Figure 4: Metal Sector: $\hat{\alpha}_{gt}$'s time-path

Figure 4 shows the homogeneous group specification understates the productivity mean level for most groups in the heterogeneous specification. The graph helps to explain the 18% difference that is documented in table 11. Appendix E has the plots for the other three sectors. Food and Textile sectors also exhibit upward growth trends. The Wood sector's productivity dynamic is more complex and requires more context for interpretation.

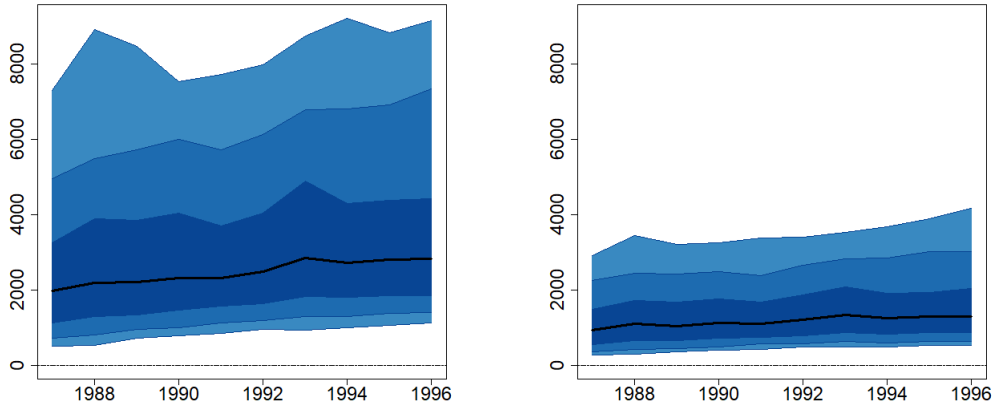


Figure 5: Metal Productivity Fan Charts: 5%,10%,25%,50%,75%,90%,95%. Left: $G = 5$ and Right: $G = 1$.

The graph shows the sectors' weighted[15] productivity distribution from 1987 to 1996. The 75 and 90 percentiles increased twice-fold after accounting for heterogeneous groups. Appendix E presents the fan charts for the other sectors, and the increase in productivity heterogeneity is also noticeable for the Food and Textile sectors.

3.6. Conclusion

This chapter build a two-step estimator to extend the proxy variable method, extensively used to estimate the firm's production function. My extension addresses

[15]The weights are the firm's market share in the sector's sample.

the proxy variable's scalar unobservable problem by introducing firms' productivity as cross-correlated. Now a firm's productivity innovation can have positive spillover effects on other firms. Furthermore, for the intermediate material's structural value-added model, the marginal product of intermediate material can now be time varying. In Monte Carlo simulation, I find my production function estimator can perform well even under a small T when the groups are well-separated. Furthermore, the information criterion can overwhelmingly select the correct number of groups under a small T.

For the empirical application, I apply my production function estimator on four large Chilean manufacturing sectors from 1987 to 1996. In line with the transmission bias intuition, my estimator downward revises the proxy variable's estimates on the more flexible input's coefficient - the output elasticity for labour. For policy analysis, my analysis shows a significant increase in the productivity distribution's dispersion after introducing heterogeneous productivity groups. Furthermore, productivity also appears more responsible for output growth in the Metal, Food, and Textile sectors.

3.7. Extensions

One natural extension is to consider the production function having a latent group structure beyond its productivity process. This extension allows latent differences among firms' production processes to determine their output elasticities. For example, firms geared towards automation are more able to substitute labour for capital. However, the degree of automation in the firm's production is heterogeneous within an industry. Moreover, knowing the firm's level of automation is not observable from the typical production data. Consequently, a model of latent groups can capture the level of automation and many other unobserved factors.

Chapter 4 provides an empirical production function model having latent groups for

both its output elasticities and productivity process. However, it models productivity as an autoregressive process. This chapter's method does not require an autoregressive assumption. Here, the extension provides a production function estimator to allow beyond-productivity latent groups while not assuming the autoregressive structure.

The key is assume the nonparametric first-stage regression m has a latent group structure, as described in Chapter 2's extension. And this group structure is the beyond-productivity latent groups. To be specific, the gross production function is

$$Y_{it} = \exp\left(\alpha^0_{g_i^0 t} + \epsilon_{it}\right) \min\{\mathcal{C}_{h_i}\left(M_{it}\right), \exp\left(\omega_{it}\right) F_{h_i}\left(K_{it}, L_{it}\right)\}, \tag{3.18}$$

where h_i denotes the beyond-productivity latent group membership. Then identifying group memberships for the first-stage's nonparametric conditional mean function also identifies groups of firms sharing the function F. Then the second-stage has a natural extension by plugging in these group memberships, h_i.

Chapter 4

Clustering for Multidimensional Heterogeneity

4.1. Introduction

Firms, individuals, and countries are heterogeneous in multiple dimensions. For example, firms can differ in their productivities, in their output elasticities of variable inputs, and in their output elasticities of capital.[1] A flexible specification of the production function ideally allows for heterogeneity in all three of these features. For practical estimation, the key question is how to specify a flexible yet parsimonious and tractable econometric model that is consistent with such multi-dimensional unobserved heterogeneity in the data. In a panel data context, this paper proposes a framework to assign multiple cluster memberships to each cross-sectional unit, where each cluster membership is determined by one particular characteristic of the unit. We estimate the memberships as well as cluster-specific and common parameters in a nonlinear generalized method of moments (GMM) framework.

Recent years have seen increasing popularity of modeling heterogeneity through clusters. In panel data analysis, allowing each cross-sectional unit to have its own regression coefficient often leads to a large number of parameters and a poor estimation of them. Instead, researchers may divide the whole population into a finite-number of

[1] Throughout the paper we will refer to these elasticities simply as variable input and capital elasticities, respectively.

clusters and explore the commonality within and differences across clusters. The cluster membership could be known (Bester and Hansen, 2016) or estimated by machine learning methods (Lin and Ng., 2012; Bai and Ando, 2016; Bonhomme and Manresa, 2015, BM hereafter; Su, Shi, and Philips, 2016, SSP hereafter). Similar to clusters, finite mixtures models can be used to model group-wise heterogeneity (Sun, 2005; Kasahara and Shimotsu, 2009). Hahn and Moon, 2010 provide economic foundations for fixed effects with a finite support. In a Bayesian setting, correlated random effects distributions modeled flexibly with Dirichlet process mixture priors can also capture forms of group heterogeneity (e.g., Liu, 2018). This paper contributes to the literature in various ways, discussed in the following paragraphs.

First, multiple clustering has the benefit of borrow strength among units that are homogeneous in one dimension but heterogeneous in other dimensions. By introducing multiple memberships, units in one cluster share some features but differ in other features. Existing methods are one-dimensional, giving only one membership to each unit and requiring units in a cluster to share all features. For example, in our empirical estimation of the production function, we pool all firms that share the same variable input elasticity together to estimate this common parameter, regardless of the other two features, i.e., productivity and capital elasticity. Yet, we allow for heterogeneity and cluster patterns in these other features. To fit the production example into the one-dimensional clustering framework, one would only assign firms to the same cluster if their production functions are identical in all dimensions. This results in much smaller cluster sizes and more cluster-specific parameters to estimate.

Second, multi-dimensional clustering is robust to sparse interactions among different features. To estimate cluster-specific parameters, we need a large number of observations from each group. The one-dimensional approach cuts the data finer by

requiring all features to be the same in a cluster, making it possible that some cluster is much smaller than others. In the context of the production function example, the one-dimensional framework requires a large number of firms with high productivity and low output elasticity. The proposed multi-dimensional approach only requires a large number of highly productive firms and a large number of firms with low output elasticity separately.

Third, we establish classification consistency of the group membership in a nonlinear GMM framework. The group membership in each dimension is estimated by the K-means method. This theoretical analysis builds on the important classification consistency result in BM. The main difference is that the group memberships here are estimated by a nonlinear GMM criterion instead of a linear least square criterion with heterogeneous intercept. We do not allow the parameters to be time-varying as in BM. To the best of our knowledge, SSP is the only paper that considered classification based on a GMM criterion. However, they restricted it to a linear IV model. Classification with other types of criteria are considered, for instance, by SSP, Liu et al., 2018, Gu and Volgushev, 2019. The asymptotic results require both large N and large T, but allow T to grow much slower than N. Thus, they are compatible with relatively short panels with a large number of cross-sectional observations. The number of clusters for each feature can be determined by a quasi-Bayesian information criterion. Homogeneity is a special case with one cluster.

Fourth, we derive the asymptotic distributions of the cluster-specific and common parameters. SSP model some parameters to be cluster-specific and some parameters to be unit-specific. The latter results in incidental parameter bias that is subsequently corrected. Different from their approach, we model the multi-dimensional heterogeneity symmetrically by assuming all heterogeneous parameters follow cluster

patterns. The added flexibility is that different parameters are associated with different memberships. Once the memberships are consistently estimated, we impose the estimated memberships and construct a pooled GMM criterion. All cluster-specific and common parameters are estimated with $\sqrt{NT}$ rate.

We use the proposed multi-dimensional clustering technique to estimate firm-level Cobb-Douglas production functions for a subset of two digit sectors defined by the North American Industry Classification System (NAICS). Within each two-digit sector, we allow for multi-dimensional group heterogeneity in terms of total factor productivity, and output elasticities with respect to variable inputs and capital. The production functions are estimated on a sequence of rolling panel data sets for publically traded firms. Using the approach of De Loecker and Warzynski, 2012, we scale the estimated variable-input elasticities by the revenue-to-variable-cost ratio to obtain an approximation of firm-level markups. We then aggregate the firm-level markups to compute an aggregate markup for each rolling sample and re-examine the rise of aggregate markups documented by De Loecker, Eeckhout, and Unger (2018). Our main finding is that the overall level of aggregate markup is lower and the rise in the markup between 1970 and 2016 is less pronounced once one accounts for group heterogeneity among publicly-traded firms within two-digit NAICS sectors.

The remainder of the paper is organized as follows. Section 4.2 describes the model and the estimation procedure. Section 4.3 provides some key regularity conditions and shows consistency of the estimators. Section 4.4 starts with some heuristic arguments on classification of group memberships with a nonlinear GMM criterion Subsequently, we provide formal results on classification consistency and the asymptotic distribution of the GMM estimator based on a pooled criterion. Section 4.5 compares the proposed multi-dimensional clustering to standard one-dimensional clustering in a Monte Carlo

simulation. The empirical analysis is presented in Section 4.6. Finally, Section 4.7 concludes. Proofs, data definitions, and additional numerical results are relegated to Appendix - Chapter 4.

Throughout the paper, we adopt the following notations. For vectors a, b, we use (a, b) to denote $(a', b')'$, unless the dimension is defined otherwise. Let $||A||$ denote the Frobenius norm of a matrix A. When A is a symmetric, let $\mu_{\max}(A)$ and $\mu_{\min}(A)$ denote the largest and smallest eigenvalues of A. Let $1\{\cdot\}$ denote the indicator function. All asymptotic results are obtain as N and T pass to infinite jointly.

4.2. Model and Estimator

We have panel data $\{w_{it} : i = 1, ..., N; t = 1, ..., T\}$ and use them to estimate unknown parameters $\theta_i = (a_i, b_i, \lambda) \in A \times B \times \Lambda$ based on moment conditions. The parameter space A, B, Λ are subsets of $R^{d_a}, R^{d_b}, R^{d_\lambda}$, respectively. To study applications where N is significantly larger than T, we provide a parsimonious model of a_i and b_i by two separate group patterns. Let $g_i \in \{1, ..., n_g\}$ denote the membership for a_i and $h_i \in \{1, ..., n_h\}$ denote the group membership for b_i. We have

$$a_i = \begin{cases} \alpha_1 & \text{if } g_i = 1 \\ \vdots & \vdots \\ \alpha_{n_g} & \text{if } g_i = n_g \end{cases} \quad \text{and } b_i = \begin{cases} \beta_1 & \text{if } h_i = 1 \\ \vdots & \vdots \\ \beta_{n_h} & \text{if } h_i = n_h \end{cases}. \tag{4.1}$$

Let

$$\alpha = (\alpha_1, ..., \alpha_{n_g}) \in R^{d_\alpha \times d_{n_g}} \text{ and } \beta = (\beta_1, ..., \beta_{n_h}) R^{d_\beta \times d_{n_h}}.$$

denote the group-specific values. We can write

$$a_i = \alpha(g_i) \text{ and } b_i = \beta(h_i), \tag{4.2}$$

where $\alpha(g_i) = \alpha_{g_i}$ denotes the g_i^{th} column of α, similarly, $\beta(h_i) = \beta_{h_i}$ denotes the h_i^{th} column of β. With the two-dimensional group patterns, the unknown parameters are

$$\theta = (\alpha, \beta, \lambda),\ G = (g_1, ..., g_N),\ H = (h_1, ..., h_N). \tag{4.3}$$

The parameter space is $(\theta, G, H) \in \overline{\Theta} \times \Gamma_G \times \Gamma_H$, where $\overline{\Theta} = A^{n_g} \times B^{n_h} \times \Lambda$ and Γ_G and Γ_H are sets of all possible partitions of $\{1, ..., N\}$ into n_g and n_h groups, respectively. We assume n_g and n_h are known for now. In practice, they can be selected by the Bayesian information criterion given below.

We assume group patterns and moment conditions hold for the true values of the parameters. For each i, let g_i^0 and h_i^0 denote the true group memberships and $\theta_i^0 = (\alpha^0(g_i^0), \beta^0(h_i^0), \lambda^0)$ denote the true value for $\theta_i = (\alpha(g_i), \beta(h_i), \lambda)$. The moment condition is

$$M_i(\theta_i^0) = E\left[m\left(w_{it}; \theta_i^0\right)\right] = 0 \tag{4.4}$$

hold for all i and t. The GMM estimator is[2]

$$\left(\hat{\theta}, \hat{G}, \hat{H}\right) = \underset{(\theta,G,H)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H}{\arg\min}\ \hat{Q}(\theta, G, H), \tag{4.5}$$

where

$$\hat{Q}(\theta, G, H) = N^{-1} \sum_{i=1}^{N} \hat{Q}_i(\theta, g_i, h_i), \tag{4.6}$$

and

$$\hat{Q}_i(\theta, g_i, h_i) = \left[T^{-1} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]' W_{iNT} \left[T^{-1} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]. \tag{4.7}$$

[2] Fernandez-Val and Lee, 2013 and SSP also use the same type of criterion in the presence of unit-specific parameters. In our case the unit-specific parameters are the group memberships.

for some finite-dimensional function $m(w_{it};\cdot) \in R^{d_m}$ and weighting matrix W_{iNT}.

Application: Production Function Estimation. Consider a Cobb-Douglas production function

$$y_{it} = a_i^0 + b_i^0 v_{it} + c_i^0 k_{it} + \omega_{it} + \varepsilon_{it}, \tag{4.8}$$

where y_{it}, k_{it}, v_{it} are the observed log output, log capital input, and log variable inputs (including labor, intermediate inputs, materials, etc), ω_{it} is an unobserved productivity shock that is known to the firm, and ε_{it} is an unobserved output shock that is realized after the factor inputs have been chosen. The productivity shock ω_{it} follows an AR(1) process

$$\omega_{it} = \rho^0 \omega_{it-1} + \xi_{it}, \tag{4.9}$$

where the innovation ξ_{it} is uncorrelated with input choices prior to period t. The output shock ε_{it} is uncorrelated with any input choices at period t and before. For a markup calculation following De Loecker and Warzynski, 2012 and De Loecker, Eeckhout, and Unger, 2018, the parameter of interest is the output elasticity of the variable input, i.e., b_i^0. The rest are nuisance parameters. As in these papers, we assume the capital input k_{it} is determined at period $t-1$ and firms choose the variable input v_{it} optimaly at period t.

Let

$$\Delta y_{it}(\rho) = y_{it} - \rho y_{it-1},\ \Delta k_{it}(\rho) = k_{it} - \rho k_{it-1},\ \Delta v_{it}(\rho) = v_{it} - \rho v_{it-1} \tag{4.10}$$

denote the differencing terms given the parameter ρ. Then we have

$$\Delta y_{it}(\rho^0) - a_i^0(1-\rho^0) - b_i^0 \Delta v_{it}(\rho^0) - c_i^0 \Delta k_{it}(\rho^0) = \xi_{it} + \left(\varepsilon_{it} - \rho^0 \varepsilon_{it-1}\right). \tag{4.11}$$

Let z_{it} denote a vector of capital and variable inputs choices prior to period t plus the constant term. In the empirical application in Section **??** we will use $z_{it} = (1, k_{it}, k_{it-1}, v_{it-1})'$. This ensures that z_{it} is uncorrelated to the right hand side of (4.11). We have the moment condition

$$E\left[z_{it}\left(\Delta y_{it}(\rho^0) - a_i^0(1-\rho^0) - b_i^0 \Delta v_{it}(\rho^0) - c_i^0 \Delta k_{it}(\rho^0)\right)\right] = 0. \tag{4.12}$$

For illustration purpose, we consider a model with two-dimensional group heterogeneity based on a_i and b_i and assume $c_i = c$ for all i. In this case, the common parameter is $\lambda = (c, \rho)$. With the two-dimensional group membership g_i and h_i for a_i and b_i, respectively, we have

$$\begin{aligned} m\left(w_{it}; \theta_i\right) &= z_{it}\left(\Delta y_{it}(\rho) - a_i(1-\rho) - b_i \Delta v_{it}(\rho) - c_i \Delta k_{it}(\rho)\right), \text{ where} \\ a_i &= \alpha(g_i),\ b_i = \beta(h_i), \text{ and } c_i = c. \end{aligned} \tag{4.13}$$

In the empirical estimation, we allow for three-dimensional heterogeneity on a_i, b_i, c_i and set the common parameter $\lambda = \rho$. In this case, each firm i has three memberships and the model can be adjusted accordingly. ▫

In practice, we compute the GMM estimator in (4.5) by Lloyd's Algorithm. Given G and H, $\widehat{\theta}$ is a GMM estimator based on $\widehat{Q}(\theta, G, H)$. Given θ and H, we minimize the GMM criterion function to determine the group memberships $\widehat{G}$. After re-estimating θ and holding G fixed, the group memberships H are also determined by the GMM criterion function. In the subsequent description of the algorithm, M is a large number that ensures that the algorithm does not terminate after one iteration and ϵ is a number close to zero that characterizes the tolerance level for improvements in the objective function.

Algorithm 1 (Lloyd's Algorithm)**.**

1. **Initialization,** $k = 0$**:** *Provide an initial guess* $(\widehat{G}^{(0)}, \widehat{H}^{(0)})$*. Let* $c = 0$ *and* $\widehat{Q}^{(0)} = M$*.*

2. **Iterations,** $s > 0$**:** *Until* $c = 1$ *execute the following steps:*

   (a) *Using the last iteration's estimate of group memberships* $(\widehat{G}^{(s-1)}, \widehat{H}^{(s-1)})$*, estimate the parameter* θ*:*

   $$\widehat{\theta} = \arg\min_{\theta \in \overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s-1)}, \widehat{H}^{(s-1)}\big).$$

   (b) *For* $i = 1, \ldots, N$*, determine the* g*-group membership:*

   $$\widehat{g}_i^{(s)} = \arg\min_{g_i \in \{1,\ldots,n_g\}} \widehat{Q}_i\big(\widehat{\theta}, g_i, \widehat{h}_i^{(s-1)}\big).$$

   (c) *Re-estimate the parameter* θ*:*

   $$\widehat{\theta}^{(s)} = \arg\min_{\theta \in \overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s)}, \widehat{H}^{(s-1)}\big).$$

   (d) *For* $i = 1, \ldots, N$*, determine the* h*-group membership:*

   $$\widehat{h}_i^{(s)} = \arg\min_{h_i \in \{1,\ldots,n_h\}} \widehat{Q}_i\big(\widehat{\theta}, \widehat{g}_i^{(s)}, h_i\big).$$

   (e) *Assess convergence: let* $\widehat{Q}^{(s)} = \widehat{Q}\big(\theta^{(s)}, \widehat{G}^{(s)}, \widehat{H}^{(s)}\big)$ *and set*

   $$c = 1\big\{\big|\widehat{Q}^{(s)} - \widehat{Q}^{(s-1)}\big| \leqslant \epsilon\big\}.$$

4.3. Assumptions and Consistent Estimation

First, we assume the following identification condition and regularity conditions on the data generating process.

Assumption ID. For any η, $\min_{1 \leqslant i \leqslant N} \inf_{||\theta_i - \theta_i^0|| > \eta} ||M_i(\theta_i)|| > \varepsilon > 0$.

Assumption R. (i) $\{w_{it}, t = 1, 2, ...\}$ are i.i.d. across i. For each i, $\{w_{it} : t = 1, 2..., \}$ is stationary strong mixing with mixing coefficients $\alpha_i(\cdot)$, where $\alpha(\cdot) = \sup_i \alpha_i(\cdot)$ satisfies $\alpha(\tau) \leqslant c_\alpha r^\tau$ for some $c_\alpha > 0$ and $r \in (0, 1)$.

(ii) The true value θ_i^0 lies in the interior of the convex compact set $\Theta = \mathcal{A} \times \mathcal{B} \times \Lambda$ for all i.

(iii) There exists a function $f(w_{it})$ such that $\sup_{\theta_i \in \Theta} ||m(w_{it}; \theta_i)|| \leqslant f(w_{it})$ and $||m(w_{it}, \theta_i) - m(w_{it}, \overline{\theta}_i)|| \leqslant f(w_{it})||\theta_i - \overline{\theta}_i||$ for all $\theta_i, \overline{\theta}_i \in \Theta$. $E|f(w_{it})|^q < \infty$ for some $q \geqslant 6$.

Application (Continued). We assume the following conditions hold for the production function estimation. (i) $\{(v_{it}, k_{it}, \xi_{it}, \varepsilon_{it}) : t = 1, ...\}$ are i.i.d. over i. For each i, $\{(v_{it}, k_{it}, \xi_{it}, \varepsilon_{it}) : t = 1, ...\}$ is stationary strong mixing that satisfies Assumption R(i). $E(\varepsilon_{it}) = 0$, $E(\xi_{it}) = 0$, $E(\varepsilon_{it} k_{it-\tau})$ for $\tau \geqslant 0$, $E(\varepsilon_{it} v_{it-\tau})$ for $\tau \geqslant 1$. (ii) $\theta_i = (a_i, b_i, c_i, \rho) \in \Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{C} \times [0, \overline{\rho}]$ for some $\overline{\rho} < 1$, where $\mathcal{A}, \mathcal{B}, \mathcal{C} \in R$ are all convext and compact. The true value θ_i^0 is in the interior of Θ. (iii) Let $x_{it}(\rho) = (1, \Delta v_{it}(\rho), \Delta k_{it}(\rho), \omega_{it-1})'$. $\mu_{\min}(E[z_{it} x_{it}(\rho)]') \geqslant \delta$ for some $\delta > 0$ for any $\rho \in [0, \overline{\rho}]$. (iv) Let $d_{it} = (1, y_{it}, y_{it-1}, v_{it}, v_{it-1}, k_{it}, k_{it-1})$. For some $C < \infty$ and $q \geqslant 6$, $E||z_{it} d_{it}||^q \leqslant C$. Assumption ID and Assumption R hold for the production function example under conditions (i)-(iv). □

Assumption NT. $N^2 = O(T^{q/2-1})$, where $q \geqslant 6$ is the constant in Assumption R1(iii).

Assumption NT allows N to be much larger than T, if the condition holds for a large q, which further translates to the moment condition in Assumption R1(iii). Alternatively, one can also impose tail condition on $f(w_{it})$ directly, as in BM.

Under Assumption R1 and NT, SSP establishes the uniform convergence result[3]

$$P\left\{\max_{1\leqslant i\leqslant N}\sup_{\theta_i\in\Theta}\left\|T^{-1}\sum_{t=1}^{T}m\left(w_{it};\theta_i\right)-E\big[m\left(w_{it};\theta_i\right)\big]\right\|\geqslant\eta\right\}=o(N^{-1}) \tag{4.14}$$

for any $\eta>0$, as $N,T\to\infty$. To establish the estimation consistency in Lemma 1 below, the convergence rate $o(N^{-1})$ can be replaced with $o(1)$ in (4.14). However, to subsequently show the K-mean classification consistency for the memberships, the $o(N^{-1})$ rate is necessary.

Assumption W. There exists nonrandom matrices W_i such that $\max_{1\leqslant i\leqslant N}\|W_{iNT}-W_i\|\to_p 0$ and $\min_{1\leqslant i\leqslant N}\mu_{\min}(W_i)=\underline{c}_W>0$ and $\max_i\mu_{\max}(W_i)=\overline{c}_W<\infty$.

Application (Continued). For the production function application, we can choose $W_{iNT}=(T^{-1}\sum_{t=1}^{T}z_{it}z_{it}')^{-1}$. It corresponds to the optimal weighting matrix if the conditional variance of the shocks are constant over time, although it may vary across i. For this choice of W_{iNT}, Assumption W holds by (4.14) and condition $E[z_iz_i']$ has full rank and $E||z_i||^2<\infty$. ▫

The following Lemma shows that the estimators are consistent on average.

Lemma 1. *Suppose Assumptions ID, R, NT, W hold. Then,*

$$N^{-1}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{g}_i)-\alpha^0(g_i^0)\right)^2\to_p 0,\ N^{-1}\sum_{i=1}^{N}\left(\widehat{\beta}(\widehat{h}_i)-\beta^0(h_i^0)\right)^2\to_p 0,\ \widehat{\lambda}\to_p\lambda^0.$$

[3]See Lemma S1.2(iii) of SSP.

Next, we consider estimation of the group specific parameters $\alpha^0 = (\alpha_1^0, ..., \alpha_{n_g}^0)$ and $\beta^0 = (\beta_1^0, ..., \beta_{n_g}^0)$. To this end, we add Assumption S, which states that each group is well separated from the rest and each group size is a non-degenerate portion of the whole population.

Assumption S. (i) For all $g \neq \widetilde{g}$, $h \neq \widetilde{h}$, $||\alpha_g^0 - \alpha_{\widetilde{g}}^0||^2 > c$ and $||\beta_h^0 - \beta_{\widetilde{h}}^0||^2 > c$ for $c > 0$.

(ii) $N^{-1} \sum_{i=1}^{n} 1\{g_i^0 = g\} \to \pi_g > 0$ and $N^{-1} \sum_{i=1}^{n} 1\{h_i^0 = h\} \to \psi_h > 0$ for all $g \in \{1, ..., n_g\}$ and $h \in \{1, ..., n_h\}$.

Assumption S(ii) allows for sparse interactions between two types, i.e.,

$$N^{-1} \sum_{i=1}^{N} 1\{g_i = g \text{ and } h_i = h\} \to 0 \quad \text{for some } (g, h).$$

One can handle the two-dimensional clustering model with the one-dimensional method by calling $\{i : g_i = g \text{ and } h_i = h\}$ a cluster. However, this one-dimensional method does not allow for sparse interactions, because the number of observations in this interaction is too small. The two-dimensional clustering method solves this problem because we estimate $\alpha(g_i)$ with all observations that share the membership g_i, regardless of h_i. The same argument holds for the estimation of $\beta(h_i)$.

Note that the criterion function $\widehat{Q}(\theta, G, H)$ is invariant to relabeling the group memberships in (θ, G, H). Without loss of generality, we assume $(\widehat{\theta}, \widehat{G}, \widehat{H})$ is already suitably relabeled such that we can show $\widehat{\alpha} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_{n_g})$ is a consistent estimator of $\alpha^0 = (\alpha_1^0, ..., \alpha_{n_g}^0)$ and $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_{n_h})$ is a consistent estimator of $\beta^0 = (\beta_1^0, ..., \beta_{n_g}^0)$ below.

Lemma 2. *Under the assumptions for Lemma 1 and Assumption S,* $\widehat{\theta} \to_p \theta^0$, *i.e.,*

$\widehat{\alpha} \to_p \alpha^0$, $\widehat{\beta} \to_p \beta^0$, $\widehat{\lambda} \to_p \lambda^0$.

It is worth pointing out that $N^{-1}\sum_{i=1}^{N}(\widehat{\alpha}(\widehat{g}_i) - \alpha^0(g_i^0))^2$ in Lemma 1 and $||\widehat{\alpha} - \alpha^0||^2$ in Lemma 2 are two different measures between the estimator and the true value. The former is based on $\widehat{\alpha}(\widehat{g}_i)$, where the group membership $\widehat{g}_i$ could be possibly misclassified. The later $\widehat{\alpha}$ does not consider the group membership classification.

4.4. Classification and Asymptotic Distribution

Given $\widehat{\theta}$, $\widehat{G}$ and $\widehat{H}$ are K-mean estimators of the group memberships that minimize the nonlinear GMM criterion function $Q(\widehat{\theta}, G, H)$. BM provide consistency of the K-mean clustering for linear least squares estimation. SSP study classification with the GMM criterion using a shrinkage procedure, but also restrict it to linear models. We extend classification consistency to nonlinear GMM problems and allow for multiple-dimensional K-mean methods.

Before presenting the formal result, we first illustrate the intuition and key arguments. For the ease of notation in subsequent arguments, write

$$m_{it}(\theta, g, h) = m(w_{it}; \alpha(g), \beta(h), \lambda), \tag{4.15}$$

for any $g \in \{1, ..., n_g\}$, $h \in \{1, ..., n_h\}$. Because $\widehat{\theta} \to_p \theta_0$, it is sufficient to consider $\widehat{\theta} \in N_\eta = \{\theta \in \Theta : ||\theta - \theta_0|| \leqslant \eta\}$ for some positive number η.

Given $\widehat{\theta}$, for any $(g_i, h_i) \neq (g_i^0, h_i^0)$, we have

$$P\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} \leqslant P\left\{\widehat{Q}_i(\widehat{\theta}, g_i, h_i) < \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0)\right\}. \tag{4.16}$$

By Assumption W,

$$\begin{aligned}\widehat{Q}_i(\widehat{\theta}, g_i, h_i) &\geqslant c_1 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2, \\ \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0) &\leqslant c_2 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \end{aligned} \tag{4.17}$$

for some positive constants c_2 and c_1, with probability approaching 1. To bound the probability of misspecifying the membership of i to (g_i, h_i), it is therefore sufficient to bound

$$P_{i,gh}(\widehat{\theta}) = P\left\{ c_1 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2 \leqslant c_2 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \right\}. \tag{4.18}$$

With a decomposition,

$$\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) = \left(\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) - E[m_{it}(\widehat{\theta}, g_i, h_i)] \right) + E[m_{it}(\widehat{\theta}, g_i, h_i)], \tag{4.19}$$

where (i) the first term on the right hand side is a $o_p(1)$ *noise* term and (ii) the second term $E[m_{it}(\widehat{\theta}, g, h)]$ is a *signal* term that is strictly positive and bounded away from 0 conditional on $\widehat{\theta} \in N_\eta$ for η small enough. This positive signal for misspecified group is ensured by the separability condition in Assumption S and the identification condition in Assumption ID. By a similar decomposition for $T^{-1}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)$, we can show that (i) the noise is also $o_p(1)$ and (ii) the signal term $E[m_{it}(\widehat{\theta}, g_i^0, h_i^0)]$ is arbitrarily small with $\widehat{\theta} \in N_\eta$ for η small enough because $E[m_{it}(\theta^0, g_i^0, h_i^0)] = 0$. We can show that, under Assumption R and NT, the probability of the noise terms being larger than the positive signal term converges to 0 at rate $o(N^{-1})$. Therefore, we have $P_{i,gh}(\widehat{\theta})$ converges to 0 at $o(N^{-1})$ rate and the who group can be classified

consistently. The result is presented in the Theorem below and its formal proof is given in the Appendix.

Theorem 5. *Suppose Assumptions ID, R, NT, W, S hold.*

$$P\left\{\widehat{G} = G^0 \text{ and } \widehat{H} = H^0\right\} \to 1 \text{ as } N, T \to \infty,$$

where $G^0 = \{g_1^0, ..., g_N^0\}$ *and* $H^0 = \{h_1^0, ..., h_N^0\}$ *are the true memberships.*

Next we study estimation of $\theta^0 = (\alpha_1^0, ..., \alpha_{n_g}^0, \beta_1^0, ..., \beta_{n_h}^0, \lambda^0)$. Given the group membership $\widehat{G}$ and $\widehat{H}$, we can estimate θ^0 by minimizing a pooled GMM criterion

$$\widetilde{\theta} = \arg\min_{\theta \in \overline{\Theta}} \widetilde{Q}(\theta), \text{ where } \widetilde{Q}(\theta) = \widetilde{m}(\theta)' W_{NT} \widetilde{m}(\theta), \tag{4.20}$$

with

$$\widetilde{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} m\left(w_{it}; \alpha(\widehat{g}_i), \beta(\widehat{h}_i), \lambda\right), \tag{4.21}$$

and W_{NT} is a weighting matrix which could depend on $\widehat{G}$ and $\widehat{H}$. In a linear instrumental variable model with heterogeneous coefficients, SSP show that the pooled estimator $\widetilde{\theta}$ is preferred to $\widehat{\theta}$ in (4.5) because $\widehat{\theta}$ typically is less efficient and suffers from asymptotic bias. Under Theorem 5, $\widetilde{\theta}$ has the same asymptotic distribution as the oracle estimator, which is defined analogous to $\widetilde{\theta}$ but imposing the true memberships G^0 and H^0. Thus, we derive the asymptotic distribution of $\widetilde{\theta}$ by studying the oracle estimator.

We first look at the first order derivative of the moment conditions. We assume that

the function $m(w_{it}, \cdot)$ is differentiable in all parameters. Define

$$m_\theta\left(w_{it};\theta_i^0\right) = \left[\frac{\partial}{\partial\alpha}m\left(w_{it};\theta_i^0\right) : \frac{\partial}{\partial\beta}m\left(w_{it};\theta_i^0\right) : \frac{\partial}{\partial\lambda}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m \times \left(d_a n_g + d_\beta n_h + d_\lambda\right)}, \tag{4.22}$$

where

$$\begin{aligned}
\frac{\partial}{\partial\alpha}m\left(w_{it};\theta_i^0\right) &= \left[\frac{\partial}{\partial\alpha_1}m\left(w_{it};\theta_i^0\right) : \cdots : \frac{\partial}{\partial\alpha_{n_g}}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m \times (d_\alpha n_g)}, \\
\frac{\partial}{\partial\beta}m\left(w_{it};\theta_i^0\right) &= \left[\frac{\partial}{\partial\beta_1}m\left(w_{it};\theta_i^0\right) : \cdots : \frac{\partial}{\partial\beta_{n_h}}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m \times \left(d_\beta n_h\right)}.
\end{aligned} \tag{4.23}$$

Under the group structure, $m(w_{it}, \theta_i^0)$ do note depend on α_g for $g \neq g_i^0$ or β_h for $h \neq h_i^0$. Thus, we have

$$\begin{aligned}
\frac{\partial}{\partial\alpha_g}m\left(w_{it};\theta_i^0\right) &= 1\left\{g_i^0 = g\right\} m_\alpha(w_{it},\theta_i^0) \text{ for } g = 1, ..., n_g, \\
\frac{\partial}{\partial\beta_h}m\left(w_{it};\theta_i^0\right) &= 1\left\{h_i^0 = h\right\} m_\beta(w_{it},\theta_i^0) \text{ for } h = 1, ..., n_h,
\end{aligned} \tag{4.24}$$

where

$$\begin{aligned}
m_\alpha(w_{it},\theta_i) &= \frac{\partial}{\partial a_i}m\left(w_{it};a_i,b_i,\lambda\right) \in R^{d_m \times d_\alpha}, \\
m_\beta(w_{it},\theta_i) &= \frac{\partial}{\partial b_i}m\left(w_{it};a_i,b_i,\lambda\right) \in R^{d_m \times d_\beta}.
\end{aligned} \tag{4.25}$$

The Jacobian matrix is

$$J = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} E\left[m_\theta\left(w_{it};\theta_i^0\right)\right]. \tag{4.26}$$

The covariance of the moment condition is

$$\begin{aligned}\Omega &= \lim_{N\to\infty}\lim_{T\to\infty} N^{-1}\sum_{i=1}^{N}\Omega_{iT}(\theta_{i,0}), \text{ where} \\ \Omega_{i_T}(\theta_i^0) &= T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left[m\left(w_{it};\theta_i^0\right) m\left(w_{is};\theta_i^0\right)'\right]. \end{aligned} \tag{4.27}$$

These limits exist because the data is strong mixing over t, i.i.d. over i, and there is a finite-number of groups whose share converges to constants. We add the following regularity condition to derive the distribution of $\widetilde{\theta}$.

Assumption E. (i) J and Ω both have full rank.

(ii) $W_{NT} \to_p W$ for some full rank matrix W as $N, T \to \infty$.

(iii) Assumption R(iii) holds with $m(w_{it};\theta_i)$ replaced by $m_\theta(w_{it};\theta_i)$ and Θ replaced by a neighborhood around θ^0.

Theorem 6. *Suppose Assumptions Suppose Assumptions ID, R, NT, W, S, E hold. Then,*

$$\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right) \to_d N(0,V), \text{ where } V = \left(J'WJ\right)^{-1} J'W\Omega W \left(J'WJ\right)^{-1}.$$

In the estimation, α_g only shows up in the moment function $m(w_{it};\alpha(\widehat{g}_i),\beta(\widehat{h}_i),\lambda)$ if $\widehat{g}_i = g$, i.e., individuals whose coefficient a_i belong to the g^{th} group. However, the estimator $\widehat{\alpha}_g$ also depends on individuals in other groups through the estimation of β and λ. This is different from the case of a one-dimensional clustering considered by linear GMM problem in SSP, where the estimator of cluster specific parameter only depends on individuals in that cluster.

Application (Continued). In this application, the Jacobian matrix is

$$J = E[m_\theta\left(w_{it};\theta_i^0\right)] = -E[z_{it}((1-\rho^0), \Delta v_{it}(\rho^0), \Delta k_{it}(\rho^0), \omega_{it-1})] \tag{4.28}$$

which is full rank under condition (iii) for this example and $\rho_0 < 1$. Let $u_{it} = \xi_{it} + \left(\varepsilon_{it} - \rho^0 \varepsilon_{it-1}\right)$. The covariance matrix is

$$\Omega = \Sigma_{j=-\infty}^{\infty} \Gamma_j, \text{ where } \Gamma_j = E[z_{it} z_{it-j}' u_{it} u_{it-j}]. \tag{4.29}$$

We assume Ω is positive definite. In the first step, we use $W_{NT} = I_{d_m}$. In the second step, we use the optimal weighting matrix $W_{NT} = \widehat{\Omega}^{-1}$, where $\widehat{\Omega}$ is a heteroskedasticity and autocorrelation consistent (HAC) covariance estimator of Ω , see Newey and West, 1987 and Andrews, 1991. In the construction of the HAC estimator, we replace the expectation with the sample average over both i and t because this is for the pooled estimator. Similarly, we can get a consistent estimator of J by replacing the expectation with the sample average over both i and t and replacing ρ^0 with the pooled estimator $\widetilde{\rho}$. Assumption E(iii) holds under condition (iv) for this example, listed below Assumption R. ▫

The GMM criterion with the optimal weighting matrix is

$$\widetilde{Q}(n_g, n_h) = \widetilde{m}(\widetilde{\theta})' \widehat{\Omega}^{-1} \widetilde{m}(\widetilde{\theta}), \tag{4.30}$$

where we make it clear that $\widetilde{m}(\theta)$ and $\widehat{\Omega}$ are constructed with classification based on n_g and n_h groups for α and β, respectively. A BIC criterion for the problem is

$$BIC(n_g, n_h) = (NT)\, \widetilde{Q}(n_g, n_h) + \log(NT)(n_g d_\alpha + n_h d_\beta). \tag{4.31}$$

In practice, we can choose (n_g, n_h) to minimize $BIC(n_g, n_h)$ with $1 \leqslant n_g \leqslant g_{\max}$ and $1 \leqslant n_h \leqslant h_{\max}$ for some user-selected upper bounds $g_{\max}$ and $h_{\max}$. Besides the BIC criterion, a wide range of penalty can be derived for model selection consistency, as shown by BM and SSP for clusters and Bai and Ng, 2002 and Cheng, L., and S., 2016 for factor models. Different from these papers, all parameter are estimated at the $\sqrt{NT}$ rate in this problem and the J statistic, i.e., $(NT)\,\widetilde{Q}(n_g, n_h)$, is a natural analog of the log-likelihood. Therefore, the BIC criterion in (4.31) is a natural choice for selecting the number of clusters. A formal testing procedure for n_g and n_h similar to that in Lu and Su, 2016 is worth investigating but is beyond the scope of this paper.

4.5. Monte Carlo Experiment

We conduct a simple location experiment to illustrate the difference between multi-dimensional and one-dimensional clustering. Let $w_{it} = (w_{1,it}, w_{2,it})'$. $\alpha(k)$, $\beta(l)$, $k, l \in \{1, 2, 3\}$ are the group parameters and the group memberships are denoted by g_i and h_i. Consequently, $\alpha(g_i)$ and $\beta(h_i)$ are our parameters of interest. We assume that the following moment condition holds at the true parameter values:

$$E\big[w_{it} - (\alpha^0(g_i^0), \beta^0(h_i^0))'\big] = 0. \tag{4.32}$$

Defining $\theta = (\alpha(1), \alpha(2), \alpha(3), \beta(1), \beta(2), \beta(3))'$ and $W_{iNT} = I$, where $I_{3\times 3}$, where $I_{3\times 3}$ identity matrix, we obtain

$$\begin{aligned} \widehat{Q}(\theta, g_i, h_i) &= \left(T^{-1}\sum_{t=1}^{T} w_{1,it} - \alpha(g_i)\right)^2 + \left(T^{-1}\sum_{t=1}^{T} w_{2,it} - \beta(h_i)\right)^2 \\ &= \big(\bar{w}_{1,i} - \alpha(g_i)\big)^2 + \big(\bar{w}_{2,i} - \beta(h_i)\big)^2, \end{aligned} \tag{4.33}$$

where $\bar{w}_{j,i}$ is the time series average of the $w_{j,it}$'s. Rather than modeling the law of motion of $w_{j,it}$ explicitly, we simply make distributional assumptions about the $\bar{w}_{j,i}$'s. For large T, we expect the sample averages to be approximately normally distributed, which is why we are assuming a data generating process (DGP) of the following form (omitting the 0 superscripts)

$$\bar{w}_i = \begin{bmatrix} \bar{w}_{1,i} \\ \bar{w}_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha(g_i) \\ \beta(h_i) \end{bmatrix}, \begin{bmatrix} \frac{1}{T} & 0 \\ 0 & \frac{1}{T} \end{bmatrix} \right), \quad g_i, h_i \in \{1, 2\}. \tag{4.34}$$

We consider the following parameterization:

$$\big[(\alpha(k), \beta(l))\big]_{k,l \in \{1,2,3\}} = \begin{bmatrix} (0.2, 0.2) & (0.2, 0.6) & (0.2, 1) \\ (0.6, 0.2) & (0.6, 0.6) & (0.6, 1) \\ (1, 0.2) & (1, 0.6) & (1, 1) \end{bmatrix}. \tag{4.35}$$

the over the grid for $T : \{5, 10, 20, 40, 60, 80, 100, 120, 160, 200, 300\}$. In addition, we consider two sample distribution designs ($N(\alpha, \beta)$ to denote the number of observations having parameter (α, β)): the balanced design as

$$\big[N(\alpha(k), \beta(l))\big]_{k,l \in \{1,2,3\}} = \begin{bmatrix} 100 & 100 & 100 \\ 100 & 100 & 100 \\ 100 & 100 & 100 \end{bmatrix}. \tag{4.36}$$

and the sparse design as

$$\big[N(\alpha(k), \beta(l))\big]_{k,l \in \{1,2,3\}} = \begin{bmatrix} 40 & 130 & 130 \\ 130 & 40 & 130 \\ 130 & 130 & 40 \end{bmatrix}. \tag{4.37}$$

The parameters $\alpha(k)$, $\beta(l)$ and the group memberships g_i and h_i are estimated based on the following objective function

$$\widehat{Q}(\theta, G, H) = N^{-1} \sum_{i=1}^{N} \big(\bar{w}_{1,i} - \alpha(g_i)\big)^2 + N^{-1} \sum_{i=1}^{N} \big(\bar{w}_{2,i} - \beta(h_i)\big)^2. \tag{4.38}$$

In our stylized DGP, the co-clustering algorithm determines the group memberships g_i from $\bar{w}_{1,i}$, whereas the group memberships h_i are determined from $\bar{w}_{2,i}$. The GMM estimator $(\widehat{\theta}, \widehat{G}, \widehat{H})$ has the following representation.

$$\widehat{\alpha}(k) = \frac{N^{-1} \sum_{i=1}^{N} \bar{w}_{1,i} 1\{\widehat{g}_i = k\}}{N^{-1} \sum_{i=1}^{N} 1\{\widehat{g}_i = k\}}, \quad \widehat{\beta}(l) = \frac{N^{-1} \sum_{i=1}^{N} \bar{w}_{2,i} 1\{\widehat{h}_i = l\}}{N^{-1} \sum_{i=1}^{N} 1\{\widehat{h}_i = l\}}, \quad k, l \in \{1, 2\}.$$

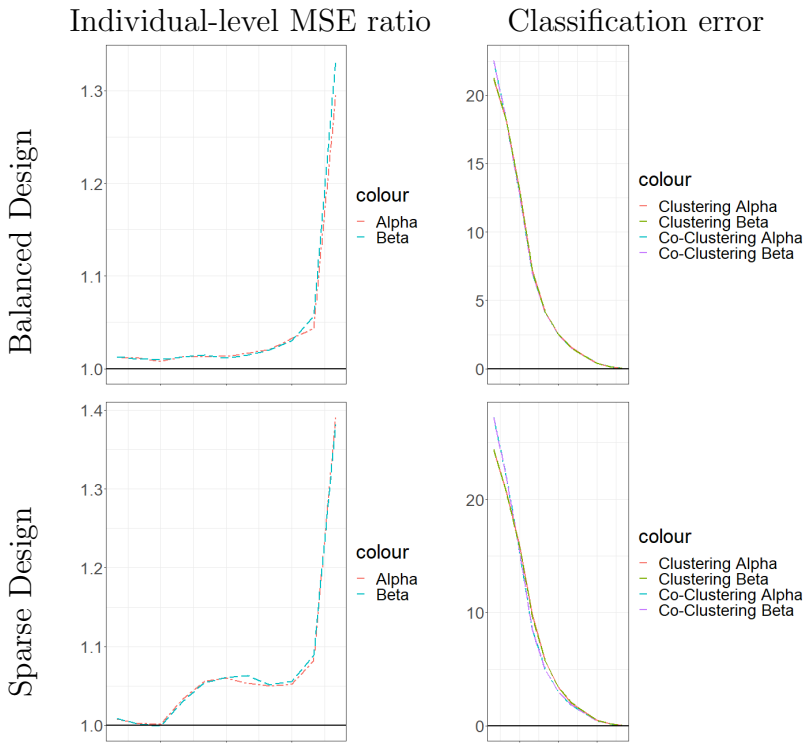
In this simple linear setting in which the estimators are sample averages, the GMM estimator $\widehat{\theta}$ is identical to pooled GMM estimator $\widetilde{\theta}$ in (4.20).

Under a single-dimensional clustering approach one would form nine separate groups which we denote by $(1,1)$, $(1,2)$, $(1,3)$, $(2,1)$, $(2,2)$, $(2,3)$, $(3,1)$, $(3,2)$, and $(3,3)$. The parameters a_i and b_i could now take on nine different values each. Accordingly, we write $a_i = \alpha_c(g_i, h_i)$ and $b_i = \beta_c(g_i, h_i)$. Here we use c subscript to indicate one-dimensional clustering. The resulting least squares objective function takes the form

$$\widehat{Q}_c(\theta_c, G, H) = N^{-1} \sum_{i=1}^{N} \big(\bar{w}_{1,i} - \alpha_c(g_i, h_i)\big)^2 + N^{-1} \sum_{i=1}^{N} \big(\bar{w}_{2,i} - \beta_c(g_i, h_i)\big)^2. \tag{4.39}$$

It is now no longer additively separable because the $\alpha_c(\cdot)$ and $\beta_c(\cdot)$ functions depend on both g_i and h_i. The standard one-dimensional clustering algorithm divides the α-β plane into nine sections.

Figure 6: Tabulated results into graphs



Notes: The left column graphs report the average ratio of single-dimensional based MSE over multi-dimensional based MSE for all observations. The right column graphs report the average classification error for all observations.

We now generate samples $n_{sim} = 1,000$ samples from the DGP in (4.34) and summarize the graphs in (4.5). We report mean-squared errors (MSE) and classification error for $(\widehat{\alpha}_i, \widehat{\beta}_i)$ and $(\hat{g}_i, \hat{h}_i)$. The estimation error for a_i can be decomposed as follows:

$$\widehat{a}_i - a_i = \big(\widehat{\alpha}(g_i) - \alpha(g_i)\big) + \big(\widehat{\alpha}(\hat{g}_i) - \widehat{\alpha}(g_i)\big).$$

The first term captures the error caused by the estimation of $\alpha(\cdot)$, assuming that the group memberships are known. In this case, all estimates are unbiased. The resulting MSEs capture the estimation variance. The one-dimensional estimator has a higher estimation variance because splitting the sample into nine groups reduces the observations per estimate. We document the two-dimensional estimator's efficiency

improvement by showing its smaller MSE in the first column graphs.

The second term captures error due to misclassification of $\hat{g}_i$. As predicted by the asymptotic theory, the classification error vanishes with a larger T - as documented by the right column graphs. Hence, the estimation error of $\alpha(\cdot)$ becomes a dominating term at the graphs' right side. Consequently, the two-dimensional estimator's efficiency improvement becomes more pronounced.

Furthermore, the two-dimensional estimator also obtains an efficiency improvement when the distribution of group memberships become less uniform. For example, in contrast to the Balanced design, three of the nine one-dimensional groups have much fewer observations in the Sparse design. This perturbation in design leads to an average precision loss for the one-dimensional estimator. On the other hand, the number of available observations per parameter is unchanged, and, consequently, the two-dimensional estimator performs similarly across the Balanced and Sparse designs. Consequently, the two-dimensional estimator experiences more MSE reduction over the one-dimensional estimator's in the Sparse design as opposed to the Balanced design. Again, this gain is evident in the left column's graphs.

It is intuitive to expect the two-dimensional estimator to obtain superior MSE performance in absence of classification error because it exploits more observations per parameter. Here, we have provided a simple Monte Carlo design showing the two-dimensional estimator to obtain superior performance even under the presence of classification error.

4.6. Empirical Analysis

Our empirical analysis re-examines the rise of aggregate markups documented by De Loecker, Eeckhout, and Unger (2018). Rising markups are a reflection of a de-

crease of competitiveness within sectors and can contribute to the observed fall of the labor share and increase in income inequality. We will show that allowing for multi-dimensional group heterogeneity within firms in two-digit NAICS sectors leads to a lower level of estimated markups and a smaller growth rate. Section 4.6.1 reviews the specification of the production function and the computation of the markups. The data set and the model specifications considered in the empirical analysis are described in Section 4.6.2. The empirical results are presented in Section 4.6.3.

4.6.1. Production Function and Markups

We will now estimate firm-level Cobb-Douglas production functions. Each firm is part of a sector d which we take to be a a two-digit NAICS sector. We follow the setup discussed in Section (**??**). The production function and the autoregressive law of motion for the unobserved productivity shock ω_{it} are given in (4.8) and (4.9), respectively. For convenience, we reproduce the equations:

$$y_{it} = a_i + b_i v_{it} + c_i k_{it} + \omega_{it} + \epsilon_{it}, \quad \omega_{it} = \rho\omega_{it-1} + \xi_{it}.$$

The GMM estimation is based on the moment conditions (4.12). Recall that the production function of is quasi-differenced to eliminate the serial correlation in ω_{it} and the vector of instruments is defined as $z_{it} = (1, k_{it}, k_{it-1}, v_{it-1})'$. We allow for group heterogeneity in a_i, b_i, and c_i. In addition to $\alpha(\cdot)$ and $\beta(\cdot)$, we define $\gamma(\cdot)$ to characterize the group-specific values of c_i. We use j_i to indicate group memberships for the third group and n_j to denote the number of groups.

Based on the estimated variable input elasticities we compute an estimate of the firms' markups. De Loecker and Warzynski (2012) show that if v_{it} induces no dynamic constraints in the firm's cost minimization problem and if the firm's capital is predetermined, then the markup can be expressed as a function of the revenue-to-

variable-cost ratio

$$mu_{it} = b_i \frac{p_{it}^y \exp[y_{it}]}{p_{it}^v \exp[v_{it}]}, \tag{4.40}$$

where p_{it}^y and p_{it}^v are firm-specific prices of the output and the variable input, respectively. Using market shares, we aggregate the firm-level markups to the sectoral level and the economy-wide level. Let $\mathcal{I}_t^d$ be the set of firms i that below to sector d, then the sector-level and the economy-wide markups are given by

$$mu_t^d = \sum_{i \in \mathcal{I}_t^d} \left(\frac{p_{it}^y \exp[y_{it}]}{\sum_{i \in \mathcal{I}_t^d} p_{it}^y \exp[y_{it}]} \right) mu_{it}, \quad mu_t = \sum_{i=1}^{N} \left(\frac{p_{it}^y \exp[y_{it}]}{\sum_{i=1}^{N} p_{it}^y \exp[y_{it}]} \right) mu_{it}. \tag{4.41}$$

4.6.2. Data Set, Model Specifications, and Estimation

As in De Loecker, Eeckhout, and Unger (2018) and Flynn, Gandhi, and Traina (2019), the firm-level data set is constructed from the Compustat Fundamentals (North America) database. We take a time period t to be one year. The firms' *Sales of Goods* and *Cost of Goods Sold* are used as output and variable input, respectively. The firms' capital stocks are calculated based on the perpetual inventory method using the *Net Property, Plant, and Equipment* series. Nominal variables are converted to real variables using the appropriate deflators. Our sample starts in 1961 and ends in 2016. Further details on data definitions, transformations, and subsample selection are provided in the Online Appendix.

There are 22 two-digit NAICS sectors. We exclude the following sectors from the subsequent analysis: Finance and Insurance (NAICS 52), Real Estate and Rental and Leasing (NAICS 53), and Public Administration (NAICS 92). Five sectors (NAICS 11, 49, 61, 71, 81) have relatively few firms so that there are not enough observations in the cross section to estimate group-specific effects. We will estimate production functions for firms in these sectors by imposing homogeneity. The 14 sectors for which

Table 12: Two-Digit-Level Sectors Used in Estimation of Models with Group Heterogeneity

NAICS	Description
21	Mining, Quarrying, and Oil and Gas Extraction
23	Construction
31	Manufacturing (Food, Apparel, and other Consumer Goods)
32	Manufacturing (Paper, Wood, Petroleum, Chemical, and Non-Metallic Minerals Related)
33	Manufacturing (Furniture, Metal, Electronic, and Machinery Related)
42	Wholesale Trade
44	Retail Trade (Food, Apparel, Vehicles, and other Consumer Goods)
45	Retail Trade (Entertainment, Department Stores, Online, etc.)
48	Transportation
51	Information
54	Professional, Scientific, and Technical Services
56	Administrative and Support Services, etc.
62	Health Care and Social Assistance
72	Accommodation and Food Services

we estimate group-specific firm-level production functions are listed in Table 12.

The subsequent analysis is conducted for firms that are associated with the same two-digit NAICS sector d. Hence, we drop the sector sub- and superscripts d if no ambiguity arises. We estimate the coefficients of the production function (4.8) for a sequence of rolling samples. The length of the rolling sample is $T = 10$ years. The first sample spans the period from 1961 to 1970 whereas the last rolling sample ranges from 2007 to 2016. The estimation for sector d includes firms for which we have at least one observation between $t = 1, \ldots, T$. We set the number of groups for a_i, b_i, and c_i equal to $n_g = n_h = n_j = 3$.[4] We refer to the results obtained from our multi-dimensional clustering estimator implemented with Algorithm 1 as *estimated heterogeneity*. In addition, we consider two alternative estimators. The *homogeneity* estimator is based on imposing that all firms within a sector d use the same production function. This

[4]There are three exceptions; see notes for Table 13.

corresponds to $n_g = n_h = n_j = 1$. The *subsector heterogeneity* estimator assumes that the production functions differ across three-digit NAICS codes. Thus, it is based on a grouping determined by a statistical agency instead of an estimation criterion.

An important set in the empirical analysis is to determine the sector-specific degree of heterogeneity in the production function coefficients. To do so, we use a quasi-Bayesian information criterion introduced in (4.31). For sample τ and model specification m, we rewrite the criterion as

$$BIC_\tau(m) = S_\tau \widetilde{Q}_{\tau,m}(\widetilde{\theta}, \widehat{G}, \widehat{H}, \widehat{J}) + k_m \log S_\tau,$$

where k_m is the number of group-specific and homogeneous coefficients and S_τ is the total number of observations in each panel τ, accounting for the fact that the panel is unbalanced.[5] Currently, we have not yet implemented a full search over $1 \leqslant n_g, n_h, n_j \leqslant \bar{n}$. Thus, we will use the criterion to compare the three above-mentioned specifications: *estimated heterogeneity*, *homogeneity*, and *subsector heterogeneity*.

4.6.3. Empirical Results

We will begin with evidence of firm heterogeneity within two-digit industries, discuss estimation results for the manufacturing sector (NAICS 32) in more detail, and then present summaries of the results across all sectors and rolling samples.

Model Selection. Table 13 summarizes the results from applying the information criteria. Rather than computing the BIC for each period separately, we are averaging over multiple samples. Columns (2) to (4) of the table contain information about the complexity, measured in terms of number of parameters, of each of the specifications. Under *estimated heterogeneity* there are generally ten free parameters: three produc-

[5]Under *subsector heterogeneity* we also estimate separate ρ's for each three-digit industry.

Table 13: Model Selection

	Complexity			Selection		
NAICS	Est.Het.	Homog.	Subsector	Est.Het.	Homog.	Subsector
21	10	4	16	X		
23	7	4	24	X		
31	10	4	24			X
32	10	4	28	X		
33	10	4	36			X
42	9	4	24			X
44	10	4	36		X	
45	10	4	16			X
48	10	4	36			X
51	10	4	40	X		
54	10	4	4	X		
56	9	4	8	X		
62	10	4	20	X		
72	10	4	8	X		

Notes: Due to data limitations we restricted the heterogeneity in three industries. For 23 we use $n_g = 3$ (productivity), $n_h = 2$ (variable inputs), $n_j = 1$ (capital). For 42 we use $n_g = 3$ (productivity), $n_h = 2$ (variable inputs), $n_j = 2$ (capital). For 56 we use $n_g = 3$ (productivity), $n_h = 3$ (variable inputs), $n_2 = 1$ (capital).

tivities $\alpha(\cdot)$, three variable input elasticities $\beta(\cdot)$, three capital elasticities $\gamma(\cdot)$, and the autoregressive coefficient ρ. For a few industries we use slightly more restrictive specifications. Under *homogeneity*, there are four parameters to estimate, and under *subsector heterogeneity* the number of parameters is four times the number of three-digit subsectors. For eight out of the fourteen sectors listed in the table, the *estimated heterogeneity* specification is preferred. For five sectors, the *subsector heterogeneity* specification attains the lowest BIC value. Because the *subsector* specification is more densely parameterized than the *estimated heterogeneity* specification, it is conceivable that the result would be overturned if we allow for additional groups.

Estimated Parameters and Groupings. Table 14 contains estimates of firm-specific productivities $\alpha(\cdot)$, variable input elasticities $\beta(\cdot)$, and capital elasticities $\gamma(\cdot)$. We consider five non-overlapping samples. Each of the samples features substantial

Table 14: 2007-2916 Parameter Estimates: Manufacturing (NAICS 32)

	Productivity			Variable Input			Capital		
Sample	$\widehat{\alpha}(1)$	$\widehat{\alpha}(2)$	$\widehat{\alpha}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\gamma}(1)$	$\widehat{\gamma}(2)$	$\widehat{\gamma}(3)$
1965-1974	.012	.015	.024	0.74	0.82	0.82	0.12	0.15	0.15
1975-1984	.045	.058	.066	0.83	0.83	0.84	0.10	0.11	0.12
1985-1994	.016	.055	.109	0.58	0.63	0.67	0.32	0.39	0.44
1995-2004	-.032	-.001	.028	0.78	0.83	0.94	0.14	0.16	0.20
2005-2014	.010	.270	.941	0.32	0.58	0.59	0.13	0.27	0.38

Table 15: Group Sizes: Manufacturing (NAICS 32), 2007-2016 Estimates, 2016 Firms

Panel (1)

	Productivity			Variable Input			Capital		
	$\widehat{\alpha}(1)$	$\widehat{\alpha}(2)$	$\widehat{\alpha}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\gamma}(1)$	$\widehat{\gamma}(2)$	$\widehat{\gamma}(3)$
Estimate	-.080	0.118	0.676	0.43	0.55	0.72	0.25	0.51	0.59
Members	180	177	5	153	47	162	142	67	153

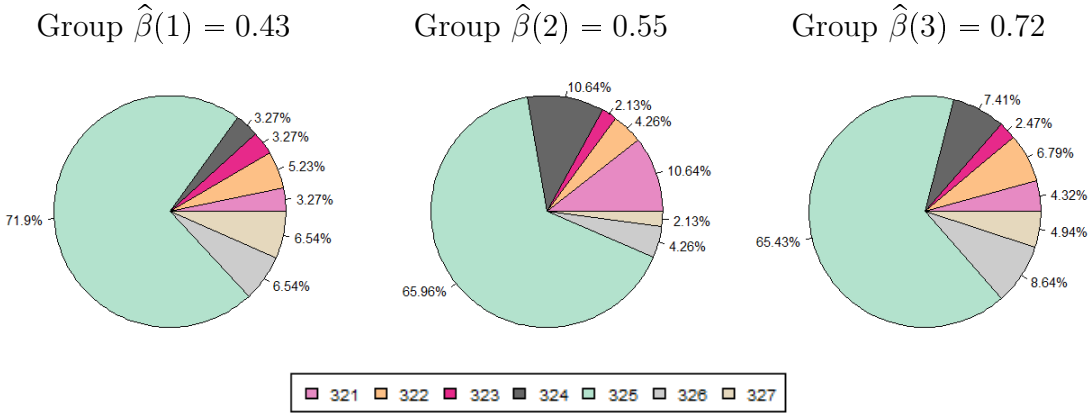
Panel (2)

	$\widehat{\alpha}(1)$			$\widehat{\alpha}(2)$			$\widehat{\alpha}(3)$		
	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$
$\widehat{\gamma}(1)$	12	1	8	89	19	11	1	0	1
$\widehat{\gamma}(2)$	3	9	27	19	0	9	0	0	0
$\widehat{\gamma}(3)$	19	15	86	9	3	18	1	0	2

heterogeneity in productivity. The heterogeneity in variable input and capital elasticities in the early samples, 1965-74 and 1975-84 is less pronounced. These periods feature only one or two, instead of three, distinct estimate of $\beta(\cdot)$ and $\gamma(\cdot)$. From 1985 onwards, the amount of heterogeneity appears to be increasing, as the parameter estimates for the three $\beta(\cdot)$ and $\gamma(\cdot)$ groups are quite different from each other.

The two panels of Table 15 provide information about the number of firms belonging to each of the groups. Here we focus on the 2007-16 sample estimates of parameters and group memberships. Because firms enter and exit the panels, we compute the number of group members for a particular year within the estimation sample,

Figure 7: Group Composition: Manufacturing (NAICS 32), 2007-2016 Estimates, 2016 Firms

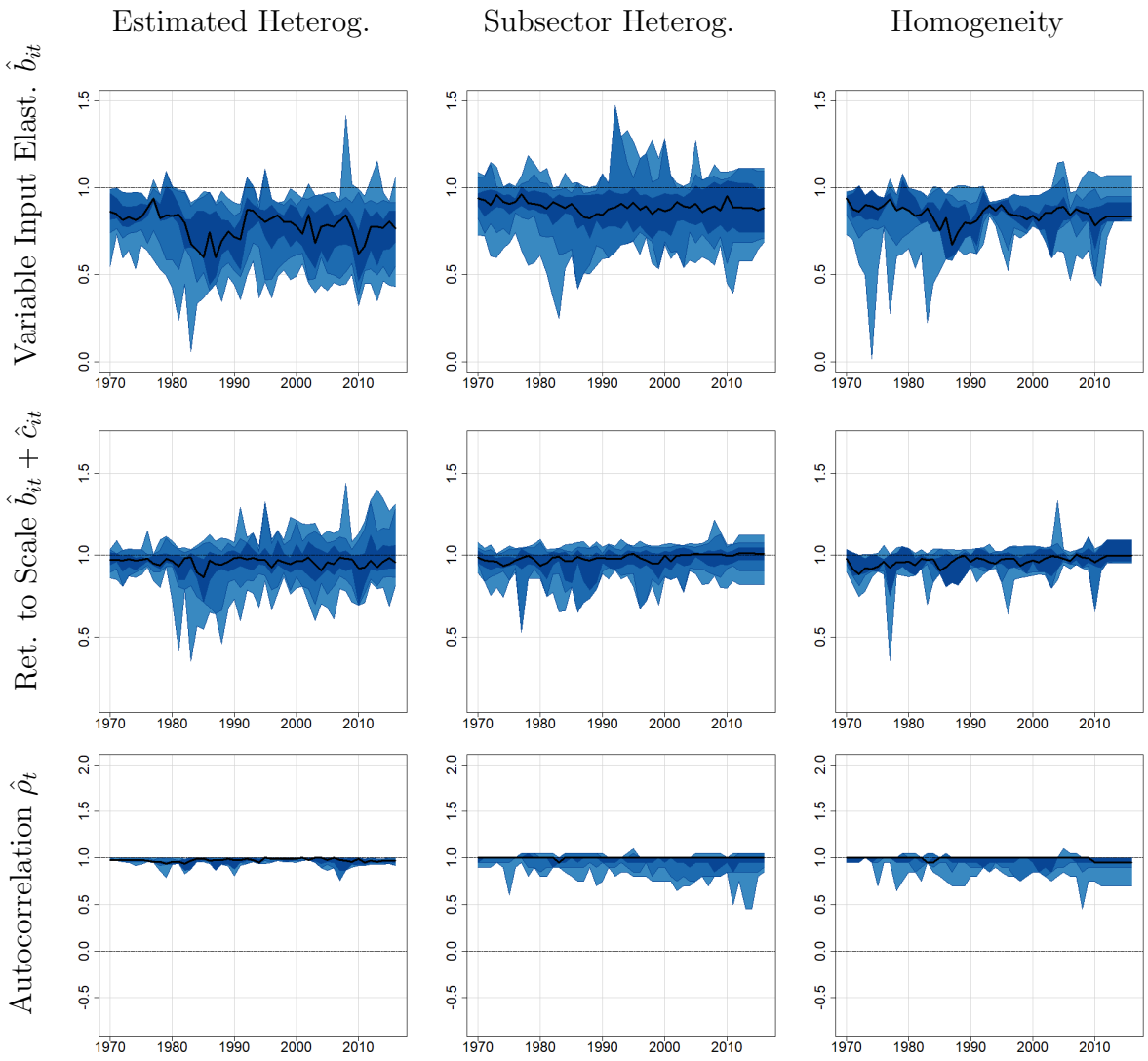


Notes: 321 = Wood Product Manufacturing, 322 = Paper Manufacturing, 323 = Printing and Related Support Activities, 324 = Petroleum and Coal Products Manufacturing, 325 = Chemical Manufacturing, 326 = Plastics and Rubber Products Manufacturing, 327 = Nonmetallic Mineral Product Manufacturing.

namely 2016. Panel (1) of the figure has the estimates of the group-specific parameters and the number of group members. Except for the high-productivity group ($\hat{\alpha}(3) = 0.676$), which only has five members and capture probably some outliers in the sample, all other groups have a substantial number of observations, allowing us to sharply estimate the group-specific coefficients. Panel (2) reports the number of firms associated with the $3 * 3 * 3 = 27$ parameter combinations that can be formed based on the nine $\alpha(\cdot)$, $\beta(\cdot)$, and $\gamma(\cdot)$ estimates. The most striking feature is that the entries in the table are sparse, in the sense that many cells have less than 10 firms. As pointed out previously, there are very few high productivity firms. More interestingly, there are few firms with medium productivity, high capital elasticity and low or medium variable input elasticity. For these sparse configurations, a one-dimensional clustering strategy based on 27 groups would have been very inefficient. Our multi-dimensional approach allows us to "extrapolate" our estimates into these sparsely-populated cells.

Figure 7 depicts the composition of the three variable cost elasticity groups for 2016.

Figure 8: Quantiles of Estimated Elasticities Across Sectors



Notes: The graphs depicts the 10%, 25%, 50%, 75%, 90%, and 95% quantiles of the cross-sectional distributions of the estimated elasticities across all two-digit sectors included in the analysis.

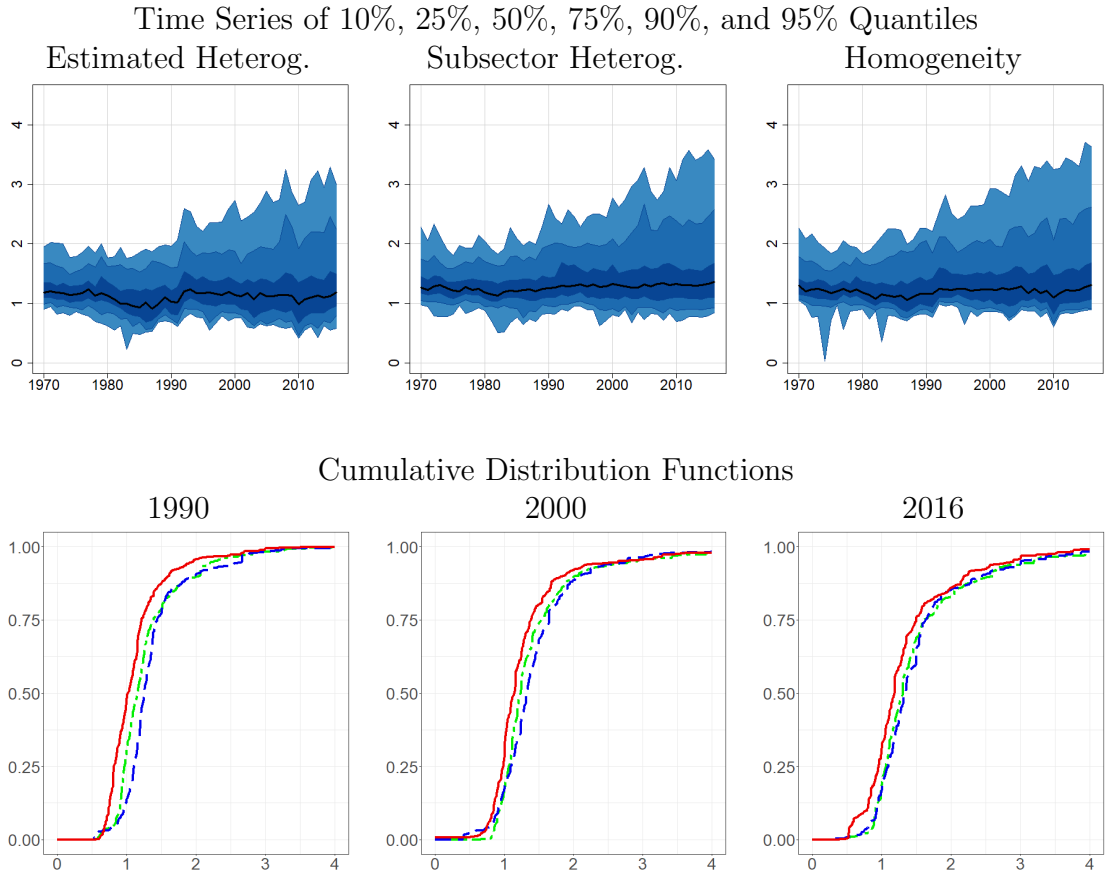
Each segment of the pie chart corresponds to a different three-digit subsector of the Manufacturing sector 32. The figure shows that each of the 7 subsectors is represented in each group. In fact, the subsector shares are very similar across $\beta(\cdot)$ groups. Thus, the estimated classification is very different from the classification of the statistical agency.

Elasticity Estimates. The firm-specific markups depend on the elasticity estimates $\hat{b}_i$ and the average markup is a function of the distribution of the $\hat{b}_i$'s within and across industries; see (4.40) and (4.41). In the top row of Figure 8 we plot quantiles of the cross-sectional distribution of the variable input elasticity estimates $\hat{b}_i$. The time series dimension of the plot traces out the sequence of rolling samples based on which we are estimating the production functions. The year on the x-axis corresponds to the midpoint (sixth observation) of each estimation sample. Because the our data set ends in 2016, the last five cross-sectional distributions for 2012 to 2016 are based on estimates from the 2007-16 sample.

The $\hat{b}_i$ estimates are weighted by the market share of firm i in that particular year. Because market shares fluctuate over time and firms enter and exit, the distribution of parameter estimates between 2012 and 2016 varies, even though the underlying estimates $\hat{\alpha}(\cdot)$, $\hat{\beta}(\cdot)$, and $\hat{\gamma}(\cdot)$ are the same. The columns of subplots in the figure correspond to the three model specifications *estimated heterogeneity*, *subsector heterogeneity*, and *homogeneity*. Under *estimated heterogeneity* the $\hat{b}_i$ estimates are lower than under *subsector heterogeneity*. By construction the estimates that impose *homogeneity* are generally less dispersed because they are identical within sector.

The second row shows the evolution of the cross-sectional distribution of the returns to scale, $\hat{b}_{it} + \hat{c}_{it}$. The sequence of medians fluctuates slightly below one indicating that the median firm operates approximately with constant returns to scale. The dispersion of the returns to scale estimates is larger under *estimated* heterogeneity than under the other two specification. This is consistent with the interpretation that grouping firms incorrectly (or imposing homogeneity), leads to estimates that average over high and low population parameters and are not representative of the dispersion in the population. The last row of Figure 8 shows the quantiles of the

Figure 9: Distribution of Markups Across Sectors

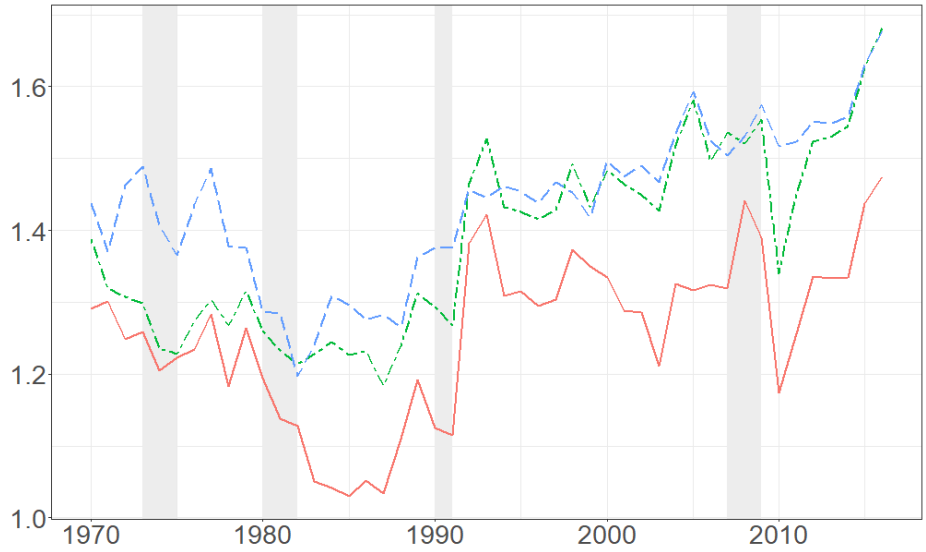


Notes: Top row: the graphs depict the evolution of the 10%, 25%, 50%, 75%, 90%, and 95% quantiles of the cross-sectional distributions of the estimated elasticities across all two-digit sectors included in the analysis. Bottom row: cumulative distribution functions for selected years based on estimated heterogeneity (red, solid), subsector heterogeneity (blue, dashed), homogeneity (green, dashed-dotted).

autocorrelation estimates. Under *estimated* heterogeneity all $\hat{\rho}$'s are very close to one, whereas for the other two specifications the estimates in the bottom quantiles often fall below 0.8.

Markup Estimates. Figure 10 shows the cross-sectional distribution of estimated markups over time, weighted by the firms' market shares. The timing convention is the same as in Figure 8. Recall that the markups are obtained by scaling the $\hat{b}_i$'s by the revenue-to-variable-cost ratio; see (4.40). Because the elasticity estimates

Figure 10: Aggregate Markups



Notes: Estimated heterogeneity (red, solid), subsector heterogeneity (blue, dashed), homogeneity (green, dashed-dotted).

obtained from the *estimated heterogeneity* specification are lower than from the other two specifications, so are the markups. In the bottom panels we show empirical distribution functions for the years 1990, 2000, and 2016. The graphs indicate a clear stochastic dominance. In all three periods, the distribution function associated with *estimated heterogeneity* lies above the distribution functions obtained from the other two specifications, indicating that the estimated markups are lower.

Using (4.41) we now compute estimates of the average markup across the sectors considered in our analysis. The results are depicted in Figure 10. The main result is that the overall level of the aggregate markup is lower and the rise in the markup between 1970 and 2016 is less pronounced under *estimated heterogeneity*, than it is under *subsector heterogeneity* and *homogeneity*. Estimated slope coefficients from a simple deterministic time trend model imply that according to the *estimated heterogeneity* version markups have risen by approximately 0.4 percentages annually. Under the *homogeneity* specification the annual increase is on average 0.7 percentages. Because

our selection criterion prefers chooses the *estimated heterogeneity* specification for the majority of sectors, we regard the resulting markup estimates from this specification as more reliable.

4.7. Conclusion

Explicitly modeling and estimating heterogeneous parameters, as opposed to simply "differencing them out" and focusing exclusively on homogeneous parameters, is an important development in the panel data literature. Our paper contributes to this literature by developing a GMM framework that allows for multi-dimensional group heterogeneity. In this framework each unit is associated with multiple groups, where each group is formed for a different characteristic of the unit. In the application, we clustered firms based on their productivity, and their elasticities of output with respect to variable inputs and capital. In our application we show that accounting for multi-dimensional group heterogeneity leads to lower estimates of the level and growth of aggregate markups than specifications that assume production technologies are homogeneous within two-digit NAICS sectors.

Bibliography

Ackerberg, D., K. Caves, and G. Frazer (2015). "Identification Properties of Recent Production Function Estimators". *Econometrica* 83 (6), pp. 2411–2451.

Ai, C. and X. Chen (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Contains Unknown Functions". *Econometrica* 71, pp. 1795–1843.

Ai, C., J. You, and Y. Zhou. (2014). ""Estimation of fixed effects panel data partially linear additive regression models"". *The Econometrics Journal* 17, pp. 83–106.

Andrews, D. (1983). "First Order Autoregressive Processes and Strong Mixing". *Cowles Foundation Discussion Paper* 664.

Andrews, D. W. K. (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation". *Econometrica* 59.3, pp. 817–858. ISSN: 00129682. DOI: 10.2307/2938229. URL: http://dx.doi.org/10.2307/2938229.

Arellano, M. (1987). "Computing Robust Standard Errors for Within-groups Estimators". *Oxford Bulletin of Economics and Statistics* 39, pp. 431–434.

Autor, D. et al. (2019). ""The Fall of the Labor Share and the Rise of Superstar Firms"".

Bai, J. (2009). "Panel Data Models with Interactive Fixed Effects". *Econometrica* 77 (4), pp. 1229–1279.

Bai, J. and T. Ando (2016). "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership". *Journal of Applied Econometrics* 31, pp. 163–191.

Bai, J. and S. Ng (2002). "Determining the Number of Factors in Approximate Factor Models". *Econometrica* 70-1, pp. 191–221.

Banerjee, A. and E. Duflo (2003). "Inequality and Growth: What Can the Data Say?" *Journal of Economic Growth* 8, pp. 267–299.

Belloni, A. et al. (2015). ""Some new asymptotic theory for least square series: pointwise and uniform results"". *Journal of Econometrics* 186, pp. 345–366.

Bester, C. Alan and Christian B. Hansen (2016). "Grouped effects estimators in fixed effects models". *Journal of Econometrics* 190.1, pp. 197–208. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2012.08.022. URL: http://www.sciencedirect.com/science/article/pii/S0304407613002030.

Bonhomme, S., T. Lamadon, and E. Manresa (2017). "Discretizing Unobserved Heterogeneity". *Working Paper*.

Bonhomme, S. and E. Manresa (2015). "Grouped Patterns of Heterogeneity in Panel Data". *Econometrica* 83 (3), pp. 1147–1184.

Chen, X. (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". *Handbook of Econometrics*. Ed. by J.J. Heckman and E.E. Leamer. 1st ed. Vol. 6B. Elsevier. Chap. 76.

Chen, X., O. Linton, and I. Keilegom (2003). "Estimation of Semiparametric Models When the Criterion Function is Not Smooth". *Econometrica* 71, pp. 1591–1608.

Cheng, X., Zhipeng L., and Frank S. (2016). "Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities". *Review of Economic Studies* 83.4, pp. 1511–1543.

Cheng, X., F. Schorfheide, and P. Shao (2019). "Clustering for Multi-dimensional Heterogeneity with Application to Production Function Estimation".

De Loecker, J., J. Eeckhout, and G. Unger (2018). "The Rise of Market Power and the Macroeconomic Implications".

De Loecker, J. and P. Scott (2017). "Estimating market power. Evidence from the US Brewing Industry".

De Loecker, J. and F. Warzynski (2012). "Markups and Firm-level Export Status". *American Economic Review* 102 (6), pp. 2437–2471.

Demirer, M. (2019). ""Production Function Estimation with Factor-Augmenting Technology: An Application to Markups"".

Doraszelski, U. and J. Jaumandreu (2018). ""Measuring the Bias of Technological Change."". *Journal of Political Economy* 126, pp. 1027–1084.

Doukhan, P. and S. Louhichi (1999). ""A new weak dependence condition and applications to moment inequalities"". *Stochastic Processes and their Applications* 84, pp. 313–342.

Fan, Jianqing. and Q.Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics.

Fernandez-Val, I. and J. Lee (2013). "Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference". *Quantitative Economics* 4.3, pp. 453–481. DOI: 10.3982/QE75. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE75. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/QE75.

Flynn, Zach, Amit Gandhi, and James Traina (2019). "Measuring Markups with Production Data". *SSRN Working Paper* 3358472.

Freyberger, J. (2018). "Non-parametric Panel Data Models with Interactive Fixed Effects". *Review of Economic Studies* 85, pp. 1824–1851.

Gandhi, A., S. Navarro, and D. Rivers (2017a). "On the Identification of Gross Output Production Functions". *University of Western Ontario, Center for Human Capital and Productivity (CHCP) Working Papers* 20181.

– (2017b). "How Heterogeneous is Productivity? A Comparison of Gross Output and Value Added". *University of Western Ontario, Center for Human Capital and Productivity (CHCP) Working Papers* 201727.

Griliches, Z. and J. Hausman (1986). ""Errors in Variables in Panel Data"". *Journal of Econometrics* 31.

Griliches, Z. and J. Mairesse (1998). "Production Functions: The Search for Identification". *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*.

Gu, J. and S. Volgushev (2019). ""Panel Data Quantile Regression with Grouped Fixed Effects"". *Journal of Econometrics* 213, pp. 68–91.

Hahn, Jinyong and Hyungsik Roger Moon (2010). "Panel Data Models with Finite Number of Multiple Equilibria". *Econometric Theory* 36.3, pp. 863–881.

Hansen, C. (2007). "Asymptotic properties of a robust variance matrix estimator for panel data when T is large". *Journal of Econometrics* 141, pp. 597–620.

Huang, X. (2013). "Nonparametric Estimation in Large Panels with Cross-sectional Dependence". *Econometric Reviews* 32 (5-6), pp. 754–777.

Imbens, G.W. and T. Lemieux. (2007). ""Regression discontinuity designs: A guide to practice"". *Journal of Econometrics*.

Kasahara, H., P. Schrimpf, and M. Suzuki (2017). "Identification and Estimation of Production Function with Unobserved Heterogeneity". *Working Paper*.

Kasahara, Hiroyuki and Katsumi Shimotsu (2009). "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices". *Econometrica* 77.1, pp. 135–175. DOI: 10.3982/ECTA6763. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6763. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6763.

Lee, J. and P. Robinson (2016). "Series estimation under cross-sectional dependence". *Journal of Econometrics* 190, pp. 1–17.

Lee, Y. (2014). ""Nonparametric Estimation of Dynamic Panel Models with Fixed Effects"". *Econometric Theory* 30, pp. 1315–1347.

Lee, Y., A. Stoyanov, and N. Zubanov (2019). "Olley and Pakes-style Producton Function Estimators with Firm Fixed Effects". *Oxford Bulletin of Economics and Statistics* 81,1, pp. 79–97.

Levinsohn, J. and A. Petrin (2003). "Estimating Production Functions Using Inputs to Control for Unobservables". *Review of Economic Studies* 70 (2), pp. 317–342.

Lin, C. and Serena. Ng. (2012). ""Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown"". *Journal of Econometric Methods* 1, pp. 42–55.

Liu, Laura (2018). "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective". *arXiv preprint 1805.04178*.

Liu, R. et al. (2018). ""Identification and estimation in panel models with overspecified number of groups"".

Lu, Xun and Liangjun Su (2016). "Determining the Number of Groups in Latent Panel Structures with an Application to Income and Democracy". *Manuscript, Singapore Management University*.

Newey, W. (1997). "Convergence Rates and Asymptotic Normality for Series Estimators". *Journal of Econometrics* 79, pp. 147–168.

Newey, Whitney K. and Kenneth D. West (1987). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". *Econometrica* 55.3, pp. 703–708.

Nickell, S. (1981). ""Biases in dynamic models with fixed effects."". *Econometrica* 6, pp. 1417–1426.

Olley, S. and A. Pakes (1995). "A Limit Theorem for a Smooth Class of Semiparametric Estimators". *Journal of Econometrics* 65, pp. 295–332.

– (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica* 64 (6), pp. 1263–1295.

Qi, L. (2000). "Efficient Estimation of Additive Partially Linear Models". *International Economic Review* 41 (4), pp. 1073–1092.

Raval, D. (2020). ""Testing the Production Approach to Markup Estimation"".

Robinson, P. (1988). "Root-N-Consistent Semiparametric Regression". *Econometrica* 56, pp. 931–954.

Su, J. and S. Jin (2012). "Sieve Estimation of Panel Data Models with Cross-section dependence". *Journal of Econometrics* 169 (1), pp. 34–47.

Su, L., Z. Shi, and P. Philips (2016). "Identifying Latent Structures in Panel Data". *Econometrica* 6, pp. 2215–2264.

Sun, Yixiao X (2005). "Estimation and Inference in Panel Structure Models". *Manuscript, University of California San Diego*. URL: `https://ideas.repec.org/p/cdl/ucsdec/qt5tf1231k.html`.

Wang, W. and L. Su. (2019). ""Identifying Latent Group Structures in Nonlinear Panels"".

APPENDIX

Appendix - Chapter 2

Extended Assumption 9:

Assumption 9.3 says, With some constant $C^{xp} > 0$, for any fixed K and $g \in \{1, ..., G^0\}$,

a there exists a sequence of stationary (jointly with $(x_{it}, z_{it}, \epsilon_{it})$) random variables, $\psi_g^x(\alpha, t) \in \mathbb{R}^{d_2}$ such that

$$\left| \psi_g^x(\alpha, t) - \frac{\sum_{i:g_i^0 = g} [x_{it} | \alpha]}{N_g} \right| \leqslant \frac{C^{xp}}{N}^1,$$

where $(x_{it}, z_{it}, \epsilon_{it}, \psi_g^x(\alpha, t))$ is jointly alpha mixing as described in Assumption 9.2.a.

b there exists a sequence of stationary (jointly with $(x_{it}, z_{it}, \epsilon_{it})$) random variables $\psi_g^m(\alpha, t) \in \mathbb{R}$ such that

$$\left| \psi_g^m(\alpha, t) - \frac{\sum_{i:g_i^0 = g} [m(z_{it}) | \alpha]}{N_g} \right| \leqslant \frac{C^{xp}}{N},$$

where $(x_{it}, z_{it}, \epsilon_{it}, \psi_g^m(\alpha, t))$ is jointly alpha mixing as described in Assumption 9.2.a.

c there exists a sequence of stationary (jointly with $(x_{it}, z_{it}, \epsilon_{it})$) random variables $\psi_g^{p,K}(\alpha, t) \in \mathbb{R}^K$ such that

$$\left| \psi_g^{p,K}(\alpha, t) - \frac{\sum_{i:g_i^0 = g} [p^K(z_{it}) | \alpha]}{N_g} \right| \leqslant \frac{C^{xp}}{N},$$

where $(x_{it}, z_{it}, \epsilon_{it}, \psi_g^{p,K}(\alpha, t))$ is jointly alpha mixing as described in Assumption 9.2.a.

[1] This upper bound can weakened to $\frac{C^{xp}}{N^{\delta_1}}$ for any $\delta_1 \geqslant \frac{1}{2}$. And, the provided proofs can be adapted to any $\delta_1 \geqslant \frac{1}{2}$ with some adjusted rates. At $\delta = \frac{1}{2}$, the conditions can be interpreted as weak laws of large numbers result. The $\delta_1 = 1$ case conveniently turns these conditions' rates as the second order issue.

d there exists a positive definite $\psi_g^{pp,K} \in \cdot^{K \times K}$ such that

$$\left| \psi_g^{pp,K} - \frac{\sum_{i:g_i^0=g} \cdot \left[\left(p^K (z_{i1}) - \psi_g^{p,K} (\alpha, 1) \right) \left(p^K (z_{i1}) - \psi_g^{p,K} (\alpha, 1) \right)' \right]}{N_g} \right| \leqslant \frac{C^{xp}}{N} \cdot$$

e there exists a positive definite $\psi^{xz} \in \cdot^{d_2 \times d_2}$ such that

$$\psi_g^{xz} = \lim_{N \to \infty} \left[\frac{\sum_{i:g_i^0=g} \cdot \left[(x_{i1} - \psi^{xx\epsilon} (z_{i1}, \alpha)) (x_{i1} - \psi^{xx\epsilon} (z_{i1}, \alpha))' \right]}{N_g} \right],$$

where $\psi^{xx\epsilon} (z_{it}, \alpha) = \cdot [x_{it} \mid z_{it}] + \psi_{g_i^0 t}^x (\alpha, t) - \cdot \left[\psi_{g_i^0 t}^x (\alpha, t) \mid z_{it} \right]$.

f the matrix $\frac{1}{N_g} \sum_{i:g_i^0=g} \cdot \left[\left(q_{i1} - \begin{pmatrix} \psi_{g_i^0}^x (\alpha, 1) \\ \psi_{g_i^0}^{p,K} (\alpha, 1) \end{pmatrix} \right) \left(q_{i1} - \begin{pmatrix} \psi_{g_i^0}^x (\alpha, 1) \\ \psi_{g_i^0}^{p,K} (\alpha, 1) \end{pmatrix} \right)' \right]$'s smallest eigenvalue is bounded below by c^{xp}.

g there exists a positive definite $\psi^{xx\epsilon} \in \cdot^{d_2 \times d_2}$ such that

$$\psi^{xx\epsilon} = \lim_{N \to \infty} \frac{\sum_{i=1}^N \left[\sum_{t=1}^\infty \cdot \left[(x_{i1} - \psi^{xx\epsilon} (z_{i1}, \alpha)) (x_{it} - \psi^{xx\epsilon} (z_{it}, \alpha))' \epsilon_{i1} \epsilon_{it} \right] \right]}{N},$$

where $\psi^{xx\epsilon} (z_{it}, \alpha) = \cdot [x_{it} \mid z_{it}] + \psi_{g_i^0 t}^x (\alpha, t) - \cdot \left[\psi_{g_i^0 t}^x (\alpha, t) \mid z_{it} \right]$.

The extended Assumption 9 adds (a), (b), (c), and (e). When the moments are identical within group,

$$\psi_g^x (\alpha, t) = \cdot \left[x_{it} | g_i^0 = g, \alpha \right],$$

$$\psi_g^m (\alpha, t) = \cdot \left[m (z_{it}) | g_i^0 = g, \alpha \right]$$

,

$$\psi_g^{p,K} (\alpha, t) = \cdot \left[p (z_{it}) | g_i^0 = g, \alpha \right],$$

$$\psi_g^{xz} = \cdot \left[(x_{i1} - \cdot [x_{i1} \mid z_{i1}]) (x_{i1} - \cdot [x_{i1} \mid z_{i1}])' | g_i^0 = g \right],$$

$$\psi^{x\epsilon} = \sum_{g=1}^{G^0} \kappa_g \sum_{t=1}^{\infty} \cdot \left[\left(x_{i1} - \cdot [x_{i1} \mid z_{i1}] \right) \left(x_{it} - \cdot [x_{it} \mid z_{it}] \right)' \epsilon_{i1} \epsilon_{it} | g_i^0 = g \right].$$

These random variables are stationary because x_{it} and z_{it} are stationary. Furthermore, if the underlying data-generating process of $(x_{it}, z_{it}, \epsilon_{it})$ is just driven by finitely many lags of $\alpha_{g_i^0 t}$ and another alpha mixing process, independent of $\alpha_{g_i^0 t}$, then the joint mixing conditions are satisfied when the conditional moments, $\psi_g^{p,K}, \psi_g^x$, and ψ_g^m, are functions of finitely many lags of α_{gt}, and α_{gt} is also an alpha mixing process.

Lemmas, Theorems, and Proofs:

Define an auxiliary criterion,

$$\bar{Q}(\theta, \beta^K, \alpha, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it}'(\theta^0 - \theta) + p^K(z_{it})'(\beta^{K,0} - \beta^K) + \alpha_{g_i^0 t} - \alpha_{g_i t})^2 + \sum_{i=1}^{N} \sum_{t=1}^{T} \epsilon_{it}^2.$$

Theorem 1 relies on $\hat{Q}$ uniformly converging to $\bar{Q}$. As Θ and $\mathcal{A}$ are compact sets, let the constant C be an upper bound for $2\|\theta\|$ and $2|\alpha_{gt}|$.

The proof of Theorem 1 uses Lemma 1 and 2.

Lemma 3. *Suppose Assumptions 1-3 hold and $\tilde{G}$ groups are used,*

$$\|\bar{Q} - \hat{Q}\|_{\infty, \Theta \times \mathcal{B}^K \times \mathcal{A}^{\tilde{G}} \times \Gamma^N_{\tilde{G}}} = O_p\left(K^{-\mu} + T^{-\frac{1}{2}} + T^{-\frac{1}{4}} + K^{\frac{1}{2}-\mu} \xi_K \Pi_K + \frac{\xi_K \Pi_K \sqrt{K}}{\sqrt{N}} \right).$$

Proof of Lemma 1. From expanding out,

$$\hat{Q}(\theta, \beta^K, \alpha, \gamma) = \bar{Q}(\theta, \beta^K, \alpha, \gamma) + A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8,$$

where

1. $A_1 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(m(z_{it}) - p^K(z_{it})' \beta^{0,K} \right) \epsilon_{it}$,
2. $A_2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(m(z_{it}) - p^K(z_{it})' \beta^{0,K} \right)^2$,

3. $A_3 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'(\theta^0-\theta)\right)\left(m\left(z_{it}\right)-p^K\left(z_{it}\right)'\beta^{0,K}\right)$,

4. $A_4 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'(\theta^0-\theta)\right)\epsilon_{it}$,

5. $A_5 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p^K\left(z_{it}\right)'\left(\beta^{0,K}-\beta\right)\right)\left(m\left(z_{it}\right)-p^K\left(z_{it}\right)'\beta^{0,K}\right)$,

6. $A_6 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p^K\left(z_{it}\right)'\left(\beta^{0,K}-\beta^K\right)\right)\epsilon_{it}$,

7. $A_7 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)\left(m\left(z_{it}\right)-p^K\left(z_{it}\right)'\beta^{0,K}\right)$, and

8. $A_8 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)\epsilon_{it}$.

On A_1 : Cauchy-Schwarz inequality (CSI), Assumption 1.3, 3.3, and 3.5 gives,

$$\begin{aligned}\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(m\left(z_{it}\right)-p^K\left(z_{it}\right)'\beta^{0,K}\right)\epsilon_{it}\right|^2\right] &\leqslant \cdot\left[\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(m\left(z_{it}\right)-p^K\left(z_{it}\right)'\beta^{0,K}\right)^2\right)\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}^2\right)\right]\\ &\leqslant \left|m-\left(p^K\right)'\beta^{0,K}\right|^2_{\infty,\mathcal{Z}}\cdot\left[\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}^2\right)\right]\\ &\leqslant O\left(K^{-2\mu}\right)\end{aligned}$$

Then Markov inequality (MI) and Jensen inequality (JI) give, $\|A_1\| = O_p(N^{-\frac{1}{2}}K^{-\mu})$.

On A_2 : Assumption 1.3 and 3.1 give $\|A_2\| = O_p(K^{-2\mu})$.

On A_3 : $\|A_3\| \leqslant \left|m-(p^K)'\beta^{0,k}\right|_{\infty,\mathcal{Z}}\left|\theta^0-\theta\right|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|x_{it}'\right|$. Then, by Assumption 1.3, 2, and 3.2, $\cdot[\|A_3\|] \leqslant C\left|m-(p^K)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}M^{\frac{3}{4}} = O_p(K^{-\mu})$. Hence, $A_3 = O_p(K^{-\mu})$.

On A_4 : JI and Assumption 3.2 give $\cdot\left[\left(\left(\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x_{it}\right|\right)^2\right] \leqslant \cdot\left[\sum_{i=1}^{N}\frac{1}{N}\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x_{it}\right|^2\right] \leqslant \frac{M}{T}$. Therefore, MI and JI implies

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x_{it} = O_p(T^{-\frac{1}{2}}).$$

Then Assumption 2 and CSI give $\|A_4\| \leqslant C\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\epsilon_{it}\right|$. Hence, $A_4 = O_p(T^{-\frac{1}{2}})$.

On A_5 : Under Assumption 1.1, 1.2, 1.3, and 2: CSI gives $\|A_5\| \leqslant 2\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}\xi_K\Pi_K\sqrt{K} = O_p\left(K^{\frac{1}{2}-\mu}\Pi_K\xi_K\right)$.

On A_6 : $\|A_6\|^2 = \frac{1}{N^2T^2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}p^K\left(z_{it}\right)'\left(\beta^{0,K}-\beta^K\right)\epsilon_{it}\right)'\left(\sum_{i=1}^{N}\sum_{t=1}^{T}p^K\left(z_{it}\right)'\left(\beta^{0,K}-\beta^K\right)\epsilon_{it}\right)$.
Then using Assumption 1.1, 1.2, 3.3, 3.5, and CSI gives,

$$\begin{aligned}\cdot\left[\|A_6\|^2\right] &= \left(\beta^{0,K}-\beta^{\beta^K}\right)'\left[\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\cdot\left[p^K\left(z_{it}\right)p^K\left(z_{js}\right)'\epsilon_{it}\epsilon_{js}\right]\right]\left(\beta^{0,K}-\beta^K\right)\\ &\leqslant \left|\beta^{0,K}-\beta^K\right|^2\left[\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\cdot\left[p^K\left(z_{it}\right)p^K\left(z_{is}\right)'\cdot\left[\epsilon_{it}\epsilon_{is}|z_{it},z_{is}\right]\right]\right]\\ &\leqslant \Pi_K^2KM^2\xi_K^2\frac{M}{N}.\end{aligned}$$

Thus MI and JI gives, $A_6 = O_p\left(\frac{\xi_K\Pi_K\sqrt{K}}{\sqrt{N}}\right)$.

On A_7 : From Assumption 1.3 and 2, $\|A_7\| \leqslant C\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}} = O_p(K^{-\mu})$.

On A_8 : Let $\{\cdot\}$ be the indicator function.

$$\cdot[\|A_8\|] \leqslant \sum_{g'=1}^{\tilde{G}}\sum_{g=1}^{G^0}\cdot\left[\left|\sum_{t=1}^{T}\left(\frac{\alpha^0_{gt}-\alpha_{g't}}{T}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\{g_i^0=g\}\{g_i=g'\}\epsilon_{it}\right)\right|\right]$$

Now CSI gives,

$$\begin{aligned}\left|\sum_{t=1}^{T}\left(\frac{\alpha_{gt}^{0}-\alpha_{g't}}{T}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\{g_i^0=g\}\{g_i=g'\}\epsilon_{it}\right)\right|^2 &\leqslant \sum_{t=1}^{T}\left(\frac{\left(\alpha_{gt}^{0}-\alpha_{g't}\right)^2}{T}\right)\left(\sum_{t=1}^{T}\frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^{N}\{g_i^0=g\}\{g_i=g'\}\epsilon_{it}\right)^2\right)\\ &\leqslant C^2\sum_{t=1}^{T}\frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^{N}\{g_i^0=g\}\{g_i=g'\}\epsilon_{it}\right)^2\\ &\leqslant C^2\sum_{t=1}^{T}\frac{1}{T}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_{it}\epsilon_{jt}\{g_i^0=g\}\{g_j^0=g\}\{g_i=g'\}\{g_j=g'\}\right)\\ &\leqslant \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it}\epsilon_{jt}-\cdot\left[\epsilon_{it}\epsilon_{jt}\right]\right)\right|\end{aligned}$$

By JI,

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it}\epsilon_{jt}-\cdot\left[\epsilon_{it}\epsilon_{jt}\right]\right)\right|\right)^2 \leqslant \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-\cdot\left[\epsilon_{it}\epsilon_{jt}\right]\right)^2.$$

Since $\cdot\left[\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-\cdot\left[\epsilon_{it}\epsilon_{jt}\right]\right)^2\right]=\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}Cov\left(\epsilon_{it}\epsilon_{jt},\epsilon_{is}\epsilon_{js}\right)$, then Assumption 3.6 bounds the above last inequality term in expectation by $\sqrt{\frac{M}{T}}$. Therefore, $A_8=O_p\left(T^{-\frac{1}{4}}\right)$.
Now collecting the terms leads to the conclusion. □

Lemma 4. *Given any $\gamma=\{g_i\}_{i=1}^N$ group assignment, its largest group $g^*\in\underset{g\in\{1,\ldots,G\}}{\arg\max}$, i.e. $g^*\in\underset{g\in\{1,..,G\}}{\arg\max}\sum_{i=1}^{N}\{\hat{g}_i=g\}$, has at least N^* elements.*

Proof of Lemma 2. This is established with a contrapositive argument.
If group g^* has less than N^* members, then at least $N-N^*$ units are assigned to the other $G-1$ groups and, in turn, the other $G-1$ groups have an average of at least $\frac{N-N^*}{G-1}$ members per group. This average is strictly than N^*. So it means one of the other group is strictly larger than group g^*. As group g^* is assumed the largest, therefore the conclusion follows. □

Proof of Theorem 1. The proof proceeds in two steps.

Step 1

The first step proves consistency of $\hat{\theta}$ and $\hat{m}$. And the second step proves the mean-squared consistency of $\hat{\alpha}$.
Notice,

$$\tilde{Q}\left(\theta,\beta^K,\alpha,\gamma\right)-\tilde{Q}\left(\theta^0,\beta^{0,K},\alpha^0,\gamma^0\right)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0-\theta\right)+p^K\left(z_{it}\right)'\left(\beta^{0,K}-\beta^K\right)+\alpha_{g_i^0t}^0-\alpha_{g_it}\right)^2.$$

And minimizing this difference with respect to $\alpha_{g_i^0t}^0-\alpha_{g_it}$ leads to a lower bound,

$$\tilde{Q}\left(\theta,\beta^K,\alpha,\gamma\right)-\tilde{Q}\left(\theta^0,\beta^{0,K},\alpha^0,\gamma^0\right)\geqslant\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix}'\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(q_{it}-\bar{q}_{g_it}\right)\left(q_{it}-\bar{q}_{g_it}\right)'\right]\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix},$$

where $\bar{q}_{gt}=\frac{\sum_{i:\hat{g}_i=g}q_{it}}{N_g}$, where $N_g=\sum_{i=1}^{N}\{\hat{g}_i=g\}$. Then,

$$\begin{aligned}\tilde{Q}\left(\theta,\beta^K,\alpha,\gamma\right)-\tilde{Q}\left(\theta^0,\beta^{0,K},\alpha^0,\gamma^0\right)&\geqslant\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix}'\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(q_{it}-\bar{q}_{g_it}\right)\left(q_{it}-\bar{q}_{g_it}\right)'\right]\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix}\\ &\geqslant\lambda_{\min}\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix}'\begin{pmatrix}\theta^0-\theta\\ \beta^{0,K}-\beta^K\end{pmatrix}\end{aligned}$$

where, $\lambda_{\min}$ is the smallest eigenvalue of $\left[\frac{1}{NTN_{g^*}^2}\sum_{i:\hat{g}_i=g^*}\sum_{t=1}^{T}\left(\sum_{j:\hat{g}_j=g^*}q_{it}-q_{jt}\right)\left(\sum_{j:\hat{g}_j=g^*}q_{it}-q_{jt}\right)'\right]$. Therefore, $\frac{1}{\lambda_{\min}}$ is the largest eigenvalue of $\left[\frac{1}{NTN_{g^*}^2}\sum_{i:\hat{g}_i=g^*}\sum_{t=1}^{T}\left(\sum_{j:\hat{g}_j=g^*}q_{it}-q_{jt}\right)\left(\sum_{j:\hat{g}_j=g^*}q_{it}-q_{jt}\right)'\right]^{-1}$. From Assumption 4 and Lemma 2, $\lambda_{\min}\geqslant\frac{1}{c}+o_p(1)$.

Hence, $\tilde{Q}\left(\theta, \beta^K, \alpha, \gamma\right) - \tilde{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right) \geqslant \frac{1}{c}\left|\begin{pmatrix}\theta^0 - \theta \\ \beta^{0,K} - \beta^K\end{pmatrix}\right|^2 + o_p(1)$.

And $\tilde{Q}\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) - \tilde{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right) = \underbrace{\tilde{Q}\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) - \hat{Q}\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right)}_{Q_1}$
$+ \underbrace{\hat{Q}\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) - \hat{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right)}_{Q_2} + \underbrace{\hat{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right) - \tilde{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right)}_{Q_3}$.

Lemma 1 implies $Q_1 = O_p(\chi)$ and $Q_3 = O_p(\chi)$, where $\chi := K^{-\mu} + T^{-\frac{1}{2}} + T^{-\frac{1}{4}} + K^{\frac{1}{2}-\mu}\xi_K \amalg_K + \frac{\xi_K \amalg_K \sqrt{K}}{\sqrt{N}}$. And $Q_3 \leqslant 0$, from definition of the estimator as the minimizer of the criterion and $G \geqslant G^0$. Therefore, $\left|\begin{pmatrix}\theta^0 - \theta \\ \beta^{0,K} - \beta^K\end{pmatrix}\right|^2 = O_p(\chi)$. And under Assumption 5, $O_p(\chi) = o_p(1)$, hence: $\left|\theta^0 - \theta\right| = o_p(1)$.
For the non-parametric estimate of $\hat{m}$,

$$\begin{aligned}\|m - \hat{m}\|_{\infty,\mathcal{Z}} &= \left|m - \left(p^K\right)'\beta^{0,K} + \left(p^K\right)'\left(\beta^{0,K} - \hat{\beta}^K\right)\right|_{\infty,\mathcal{Z}} \\ &\leqslant \left|m - \left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}} + \left|\left(p^K\right)'\left(\beta^{0,K} - \hat{\beta}^K\right)\right|_{\infty,\mathcal{Z}} \\ &\leqslant O_p\left(K^{-\mu}\right) + \xi_K\sqrt{O_p(\chi)}\end{aligned}$$

Then Assumption 5 leads to $\xi_K^2 \chi \underset{N,T,K\to\infty}{\to} 0$. Hence, $\|m - \hat{m}\|_{\infty,\mathcal{Z}} = o_p(1)$.

Step 2

By expanding out,

$$\tilde{Q}\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) - \tilde{Q}\left(\theta^0, \beta^{0,K}, \alpha^0, \gamma^0\right) = V_1 + V_2 + V_3 + V_4 + V_5 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha^0_{g_i^0 t} - \hat{\alpha}_{\hat{g}_i t}\right)^2,$$

where

1. $V_1 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0 - \hat{\theta}\right)\right)^2$,
2. $V_2 = 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0 - \hat{\theta}\right)\right)\left(p^K(z_{it})'\left(\beta^{0,K} - \hat{\beta}^K\right)\right)$,
3. $V_3 = 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0 - \hat{\theta}\right)\right)\left(\alpha^0_{g_i^0 t} - \hat{\alpha}_{\hat{g}_i t}\right)$,
4. $V_4 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p^K(z_{it})'\left(\beta^{0,K} - \hat{\beta}^K\right)\right)^2$, and
5. $V_5 = 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p^K(z_{it})'\left(\beta^{0,K} - \hat{\beta}^K\right)\right)\left(\alpha^0_{g_i^0 t} - \hat{\alpha}_{\hat{g}_i t}\right)$.

From Assumption 5, $\left|\theta^0 - \hat{\theta}\right|, \left|\beta^{0,K} - \hat{\beta}^K\right|, \xi_K\left|\beta^{0,K} - \hat{\beta}^K\right| \xrightarrow{P} 0$, as $N, T, K \to \infty$. This observation is repeatedly applied below.
On V_1: From Assumption 3.2,

$$\begin{aligned}\cdot[V_1] &\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\cdot\left[|x_{it}|^2\left|\theta^0 - \hat{\theta}\right|^2\right] \\ &\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sqrt{\cdot\left[|x_{it}|^4\right]}\sqrt{\left[\left|\theta^0 - \hat{\theta}\right|^4\right]} \\ &\leqslant \sqrt{M}\sqrt{\left[\left|\theta^0 - \hat{\theta}\right|^4\right]} = o_p(1)\end{aligned}$$

On V_2: From Assumption 1.1 and 3.2,

$$\begin{aligned}\mathbb{E}[\|V_2\|] &\leqslant 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[|x_{it}|\left|\theta^0-\hat{\theta}\right|\left|p^K(z_{it})\right|\left|\beta^{0,K}-\hat{\beta}^K\right|\right]\\ &\leqslant 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[|x_{it}|^2\right]}\sqrt{\mathbb{E}\left[\xi_K^2\left|\beta^{0,K}-\hat{\beta}^K\right|^2\left|\theta^0-\hat{\theta}\right|^2\right]}\\ &\leqslant 2\sqrt{M}\sqrt{\mathbb{E}\left[\xi_K^2\left|\beta^{0,K}-\hat{\beta}^K\right|^2\left|\theta^0-\hat{\theta}\right|^2\right]}=o_p(1)\end{aligned}$$

On V_3: From Assumption 2 and 3.2,

$$\begin{aligned}\mathbb{E}[\|V_3\|] &\leqslant 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[|x_{it}|\left|\theta^0-\hat{\theta}\right|\left|\alpha^0_{g_i^0 t}-\hat{\alpha}_{\hat{g}_i t}\right|\right]\\ &\leqslant 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[|x_{it}|^2\right]}\sqrt{\mathbb{E}\left[\left|\theta^0-\hat{\theta}\right|^2\left|\alpha^0_{g_i^0 t}-\hat{\alpha}_{\hat{g}_i t}\right|^2\right]}\\ &\leqslant 2\sqrt{M}C\sqrt{\mathbb{E}\left[\left|\theta^0-\hat{\theta}\right|^2\right]}=o_p(1)\end{aligned}$$

On V_4: From Assumption 1.1,

$$\begin{aligned}\mathbb{E}[\|V_4\|] &\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\left|p^K(z_{it})\right|^2\left|\beta^{0,K}-\hat{\beta}^K\right|^2\right]\\ &\leqslant 2\left[\xi_K^2\left|\beta^{0,K}-\hat{\beta}^K\right|^2\right]\end{aligned}$$

On V_5: From Assumption 1.1 and 2,

$$\begin{aligned}\mathbb{E}[\|V_5\|] &\leqslant 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\left|p^K(z_{it})\right|\left|\beta^{0,K}-\hat{\beta}^K\right|\left|\alpha^0_{g_i^0 t}-\hat{\alpha}_{\hat{g}_i t}\right|\right]\\ &\leqslant 2C\mathbb{E}\left[\xi_K\left|\beta^{0,K}-\hat{\beta}^K\right|\right]=o_p(1)\end{aligned}$$

Under Assumption 5, Step 1 shows $\tilde{Q}\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha},\hat{\gamma}\right)-\tilde{Q}\left(\theta^0,\beta^{0,K},\alpha^0,\gamma^0\right)=o_p(1)$ and, thus, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha^0_{g_i^0 t}-\hat{\alpha}_{\hat{g}_i t}\right)^2=o_p(1)$. This will be used in Corollary 1. □

Next, proof of Corollary 1 follows.

Proof of Corollary 1. The argument is similar to the proof of Bonhomme and Manresa (2015)'s Lemma B.3. For completeness, the paper presents it here.

Define $\phi\left(g^0\right)\in\underset{g\in\{1,\ldots,G\}}{\arg\min}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g^0 t}-\hat{\alpha}_{gt}\right)^2$. The objective is to show,

1. $\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g^0 t}-\hat{\alpha}_{\phi\left(g^0\right)t}\right)\xrightarrow{P}0$, and
2. ϕ is bijective when $G^0=G$.

On 1: From the definition of ϕ as a minimum and $N_{g^0} = \sum_{i=1}^{N} \{g_i^0 = g^0\}$,

$$\begin{aligned}\frac{N_g}{N}\frac{1}{T}\sum_{i=1}^{T}\left(\hat{\alpha}_{\phi\left(g^0\right)t} - \alpha^0_{g^0 t}\right)^2 &\leqslant \frac{1}{N}\sum_{i=1}^{N}\{g_i^0 = g\}\frac{1}{T}\sum_{i=1}^{T}\left(\hat{\alpha}_{\phi\left(g^0\right)t} - \alpha^0_{g^0 t}\right)^2 \\ &\leqslant \frac{1}{N}\sum_{i=1}^{N}\{g_i^0 = g\}\frac{1}{T}\sum_{i=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g^0 t}\right)^2 \\ &\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g^0 t}\right)^2 = o_p(1).\end{aligned}$$

The last equality comes from Theorem 1's proof. Then, by Assumption 6.2, $\frac{1}{T}\sum_{i=1}^{T}\left(\hat{\alpha}_{\phi\left(g^0\right)t} - \alpha^0_{g^0 t}\right)^2 \leqslant \left(\frac{1}{c} + o_p(1)\right) o_p(1)$. **On 2:** Let $g_i^0 \neq g_j^0$. Then by triangle inequality,

$$\begin{aligned}&\left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\phi\left(g_i^0\right)t} - \hat{\alpha}_{\phi\left(g_j^0\right)t}\right)^2\right)^{\frac{1}{2}} + \left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\phi\left(g_i^0\right)t} - \alpha^0_{g_j^0 t}\right)^2\right)^{\frac{1}{2}} + \left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\phi\left(g_j^0\right)t} - \alpha^0_{g_j^0 t}\right)^2\right)^{\frac{1}{2}} \\ &\geqslant \left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g_i^0 t} - \alpha^0_{g_j^0 t}\right)^2\right)^{\frac{1}{2}}\end{aligned}$$

The LHS's last two terms are $o_p(1)$ from 1 and the RHS's limit is bounded below by $c^{\frac{1}{2}}$, from Assumption 6.1.a. Thus, $\phi\left(g_i^0\right) \neq \phi\left(g_j^0\right)$, makes ϕ as injective.
Now since ϕ is injective and $G^0 = G$, then ϕ must be surjective. □

Next, the proof of Theorem 2 follows.

Proof of Theorem 2. The proof consists of two cases: 1) $G < G^0$ and 2) $G > G^0$.
This identity holds:

$$\hat{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}^K_{G^0}, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right) - \hat{Q}\left(\hat{\theta}_G, \hat{\beta}^K_G, \hat{\alpha}_G, \hat{\gamma}_G\right) = P_1 + P_2 + \tilde{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}^K_{G^0}, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right) - \tilde{Q}\left(\hat{\theta}_G, \hat{\beta}^K_G, \hat{\alpha}_G, \hat{\gamma}_G\right),$$

where $P_1 = \hat{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}^K_{G^0}, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right) - \tilde{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}^K_{G^0}, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right)$ and $P_2 = \tilde{Q}\left(\hat{\theta}_G, \hat{\beta}^K_G, \hat{\alpha}_G, \hat{\gamma}_G\right) - \hat{Q}\left(\hat{\theta}_G, \hat{\beta}^K_G, \hat{\alpha}_G, \hat{\gamma}_G\right)$.
Recall from Step 1 of Theorem 1's proof,

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x'_{it}\left(\theta^0 - \hat{\theta}_G\right) + p^K(z_{it})'\left(\beta^{0,K} - \hat{\beta}^K_G\right) + \alpha^0_{g_i^0 t} - \hat{\alpha}^G_{\hat{g}_i G t}\right)^2 \\ &= O_p\left(K^{-\mu} + T^{-\frac{1}{2}} + T^{-\frac{1}{4}} + K^{\frac{1}{2}-\mu}\xi_K \Pi_K + \frac{\xi_K \Pi_K \sqrt{K}}{\sqrt{N}}\right)\end{aligned}$$

whenever $G \geqslant G^0$. This fact is referenced for use in both cases.
On case 1:
Since $G < G^0$, then ϕ is not injective. Hence, there exist $g_m^0, g_l^0 \in \{1, \ldots, G^0\}$ such that $\phi\left(g_m^0\right) = \phi\left(g_l^0\right)$, but $g_m^0 \neq g_l^0$.
Then notice by triangle inequality,

$$\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g_m^0 t} - \hat{\alpha}_{\phi(g_m^0)t}\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g_l^0 t} - \hat{\alpha}_{\phi(g_l^0)t}\right)^2 \geqslant \frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g_m^0 t} - \alpha^0_{g_l^0 t}\right)^2$$

Without loss generality, Assumption 6.1 implies $\operatorname{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{g_m^0 t} - \hat{\alpha}_{\phi(g_m^0)t}\right)^2 > \frac{1}{2}c$. Now let $g_m^* \in \underset{g\in\{1,\ldots,G\}}{\arg\max}\sum_{i=1}^{N}\{g_i^0 = g_m^0 \text{ and } \hat{g}_i = g\}$ and $S_m := \{i \mid g_i^0 = g_m^0 \text{ and } \hat{g}_i = g_m^*\}$.
Under Assumption 5, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x'_{it}\left(\theta^0 - \hat{\theta}_{G^0}\right) + p^K(z_{it})'\left(\beta^{0,K} - \hat{\beta}^K_{G^0}\right) + \alpha^0_{g_i^0 t} - \hat{\alpha}^{G^0}_{\hat{g}_i G^0 t}\right)^2 = o_p(1)$.

With $N_S = \sum_{i=1}^{N} \{i \in S_m\}$,

$$
\begin{aligned}
&\bar{Q}\left(\hat{\theta}_G, \hat{\beta}_G^K, \hat{\alpha}_G, \hat{\gamma}_G\right) - \bar{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}_{G^0}^K, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right) \\
&= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it}' \left(\theta^0 - \hat{\theta}_G\right) + p^K\left(z_{it}\right)' \left(\beta^{0,K} - \hat{\beta}_G^K\right) + \alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i^G t}^G\right)^2 + o_p(1) \\
&\geqslant \frac{1}{NT} \sum_{i \in S_m} \sum_{t=1}^{T} \left(x_{it}' \left(\theta^0 - \hat{\theta}_G\right) + p^K\left(z_{it}\right)' \left(\beta^{0,K} - \hat{\beta}_G^K\right) + \alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i^G t}^G\right)^2 + o_p(1) \\
&= \frac{N_S}{N} \frac{1}{T} \sum_{t=1}^{T} \left(\alpha_{g_m^0 t}^0 - \hat{\alpha}_{\hat{g}_m t}\right)^2 \\
&- \frac{N_S}{N} \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} \left(\alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i t}\right) q_{it}'\right) \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} q_{it} q_{it}'\right)^{-1} \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} q_{it} \left(\alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i t}\right)\right) \\
&+ \frac{N_S}{N} \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} w_{it} q_{it}'\right) \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} q_{it} q_{it}'\right)^{-1} \left(\frac{1}{N_S T} \sum_{i \in S_m} \sum_{t=1}^{T} q_{it} w_{it}\right),
\end{aligned}
$$

where $w_{it} = \left(\alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i t} + q_{it}' \begin{pmatrix} \theta^0 - \hat{\theta} \\ \beta^{0,K} - \hat{\beta} \end{pmatrix}\right)$.

The third term is always non-negative and the difference of the two first terms has a positive probability limit, by Assumption 6.1.b and 6.2. Hence, $\underset{N,T,K \to \infty}{\text{plim}} \left[\bar{Q}\left(\hat{\theta}_G, \hat{\beta}_G^K, \hat{\alpha}_G, \hat{\gamma}_G\right) - \bar{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}_{G^0}^K, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right)\right] > 0$. As $P_1 = o_p(1)$, $P_2 = o_p(1)$, $\nu_T = o_p(1)$ (Assumption 7.1), therefore

$$
IC(G) - IC\left(G^0\right) = o_p(1) + \bar{Q}\left(\hat{\theta}_G, \hat{\beta}_G^K, \hat{\alpha}_G, \hat{\gamma}_G\right) - \bar{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}_{G^0}^K, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right),
$$

and, therefore, $\underset{N,K,T \to \infty}{\text{plim}} \cdot \left(IC(G) > IC\left(G^0\right)\right) = 1$ when $G < G^0$.

On case 2:
Under Assumption 7,

$$
T^{\frac{1}{4}} O_p\left(K^{-\mu} + T^{-\frac{1}{2}} + T^{-\frac{1}{4}} + K^{\frac{1}{2}-\mu} \xi_K \amalg_K + \frac{\xi_K \amalg_K \sqrt{K}}{\sqrt{N}}\right) = O_p(1).
$$

Hence, when $G > G^0$,

$$
\begin{aligned}
T^{\frac{1}{4}} \left(IC(G) - IC\left(G^0\right)\right) &= T^{\frac{1}{4}} \left(\bar{Q}\left(\hat{\theta}_G, \hat{\beta}_G^K, \hat{\alpha}_G, \hat{\gamma}_G\right) - \bar{Q}\left(\hat{\theta}_{G^0}, \hat{\beta}_{G^0}^K, \hat{\alpha}_{G^0}, \hat{\gamma}_{G^0}\right)\right) + T^{\frac{1}{4}} \nu_T \left(G - G^0\right) \\
&= O_p(1) + T^{\frac{1}{4}} \nu_T \left(G - G^0\right)
\end{aligned}
$$

By Assumption 7.2, $T^{\frac{1}{4}} \nu_T \left(G - G^0\right) \to \infty$. Thus $\underset{N,K,T \to \infty}{\text{plim}} \cdot \left(IC(G) > IC\left(G^0\right)\right) = 1$, when $G > G^0$. **Conclusion:**

$$
\begin{aligned}
\cdot \left(\hat{G^0} = G^0\right) &= \cdot \left(\bigcap_{G \in \{\underline{G}, \overline{G}\} \setminus \{G^0\}} \{IC(G) > IC\left(G^0\right)\}\right) \\
&= 1 - \cdot \left(\bigcup_{G \in \{\underline{G}, \overline{G}\} \setminus \{G^0\}} \{IC(G) \leqslant IC\left(G^0\right)\}\right) \\
&\geqslant 1 - \sum_{G \in \{\underline{G}, \overline{G}\} \setminus \{G^0\}} \left(1 - \cdot \left(IC(G) > IC\left(G^0\right)\right)\right)
\end{aligned}
$$

The last line follows from Boole's inequality. Then the above two cases imply, $\underset{N,T,K \to \infty}{\text{plim}} \cdot \left(\hat{G^0} = G^0\right) = 1$. □

The proof of Theorem 3 uses Lemma 3, 4, and 5.

Lemma 5. *Under Assumption 2, there exists a $C > 0$ such that if,*

$$
\sum_{t=1}^{T} \left(y_{it} - x_{it}'(\theta) - p_K\left(z_{it}\right)' \beta^K - \alpha_{gt}\right) \leqslant \sum_{t=1}^{T} \left(y_{it} - x_{it}'(\theta) - p_K\left(z_{it}\right)' \beta^K - \alpha_{\bar{g}t}\right),
$$

then,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{\tilde{g}t}^{0}\right)\epsilon_{it} \leqslant & -\frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{\tilde{g}t}^{0}\right)^{2}+\frac{\sqrt{8}}{2}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}} \\ & +\frac{\sqrt{8}}{2}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2}\right)^{\frac{1}{2}}+C\left|\theta^{0}-\theta\right|\left(\frac{1}{T}\sum_{t=1}^{T}\left|x_{it}\right|^{2}\right)^{\frac{1}{2}}+C\xi_{K}\left|\beta^{0,K}-\beta^{K}\right| \\ & +C\left|m-\left(p^{K}\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}+\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}} \\ & +\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2}\right)^{\frac{1}{2}}.\end{aligned}$$

Proof. $\mathcal{A}$ is compact. Therefore, there exists a C leading to $2|a| \leqslant C$ for any $a \in \mathcal{A}$.
$\sum_{t=1}^{T}\left(y_{it}-x_{it}'(\theta)-p^{K}(z_{it})'\beta^{K}-\alpha_{gt}\right)^{2} \leqslant \sum_{t=1}^{T}\left(y_{it}-x_{it}'(\theta)-p^{K}(z_{it})'\beta^{K}-\alpha_{\tilde{g}t}\right)^{2}$ implies

$$0 \leqslant \sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)\left(x_{it}'\left(\theta^{0}-\theta\right)+p^{K}(z_{it})'\left(\beta^{0,K}-\beta^{K}\right)+m(z_{it})-p^{K}(z_{it})'\beta^{0,K}+\epsilon_{it}\right)-\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{\tilde{g}t}\right)^{2} \quad (*).$$

By Cauchy-Schwarz inequality,

1.

$$\begin{aligned}\left|\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)x_{it}'\left(\theta^{0}-\theta\right)\right| &\leqslant T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^{0}-\theta\right)\right)^{2}\right)^{\frac{1}{2}} \\ &\leqslant TC\left(\frac{1}{T}\sum_{t=1}^{T}\left|x_{it}\right|^{2}\right)^{\frac{1}{2}}\left|\theta^{0}-\theta\right|\end{aligned}$$

2.

$$\begin{aligned}\left|\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)p^{K}(z_{it})'\left(\beta^{0,K}-\beta^{K}\right)\right| &\leqslant T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(p^{K}(z_{it})'\left(\beta^{0,K}-\beta^{K}\right)\right)^{2}\right)^{\frac{1}{2}} \\ &\leqslant CT\xi_{K}\left|\beta^{0,K}-\beta^{K}\right|\end{aligned}$$

3.

$$\begin{aligned}\left|\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)\left(m(z_{it})-p^{K}(z_{it})'\beta^{K}\right)\right| &\leqslant T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(m(z_{it})-p^{K}(z_{it})'\beta^{0,K}\right)^{2}\right)^{\frac{1}{2}} \\ &\leqslant TC\left|m-\left(p^{K}\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}\end{aligned}$$

4.

$$\begin{aligned}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{gt}\right)\epsilon_{it} &\leqslant \sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}^{0}-\alpha_{gt}^{0}\right)\epsilon_{it}+\left|\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{gt}\right)\epsilon_{it}\right|+\left|\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)\epsilon_{it}\right| \\ &\leqslant \sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}^{0}-\alpha_{gt}^{0}\right)\epsilon_{it}+T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{gt}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^{2}\right)^{\frac{1}{2}} \\ &\quad +T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}^{0}-\alpha_{\tilde{g}t}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^{2}\right)^{\frac{1}{2}}\end{aligned}$$

5.

$$\begin{aligned}\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{\tilde{g}t}^{0}\right)^{2}-\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{\tilde{g}t}\right)^{2} &= \sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)\left(\alpha_{gt}+\alpha_{gt}^{0}-\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)\\ &+\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)\left(\alpha_{gt}+\alpha_{gt}^{0}-\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)\\ &\leqslant T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}+\alpha_{gt}^{0}-\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}}\\ &+T\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}+\alpha_{gt}^{0}-\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}}\\ &\leqslant T\sqrt{8}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2}\right)^{\frac{1}{2}}+T\sqrt{8}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}}\end{aligned}$$

Hence,

$$-\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{\tilde{g}t}\right)^{2}\leqslant-\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{gt}^{0}-\alpha_{\tilde{g}t}^{0}\right)^{2}+T\frac{\sqrt{8}}{2}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2}\right)^{\frac{1}{2}}+T\frac{\sqrt{8}}{2}C\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\tilde{g}t}-\alpha_{\tilde{g}t}^{0}\right)^{2}\right)^{\frac{1}{2}}$$

.

Applying all these upper bounds to $(*)$ delivers the conclusion. □

Lemma 6. *Let v_t be a strongly mixing process with zero mean with mixing coefficient $\rho(t)$. If there exists constants $b_1, b_2, b_3, b_4 > 0$ such that $\rho(t) \leqslant e^{-b_1 t^{b_2}}$ and tail bound $(|v_t| > v) \leqslant e^{1-\left(\frac{v}{b_3}\right)^{b_4}}$ for any v, then there exists a function $f(T, z, b_1, b_2, b_3, b_4)$ such that*

1. $\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geqslant z\right) \leqslant f(T, z, b_1, b_2, b_3, b_4)$, *and*
2. $T^{\delta}\kappa(T, b_1, b_2, b_3, b_4) \to 0$, *as $T \to \infty$, for any $\delta > 0$.*

Proof. This is exactly Bonhomme and Manresa (2015)'s Lemma B.5. More specifically, the upper bound of $\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geqslant z\right)$ □

Let $B(\eta, K, T) := \left\{\left(\theta, \beta^{K}, \alpha\right) \middle| \left|\theta-\theta^{0}\right|^{2} \leqslant \eta, \left|\beta^{K}-\beta^{0,K}\right|^{2} \leqslant \eta, \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^{0}\right)^{2} \leqslant \eta, \text{ and } \xi_{K}^{2}\left|\beta^{K}-\beta^{0,K}\right|^{2} \leqslant \eta\right\}$

and

$$\hat{g}\left(\theta, \beta^{K}, \alpha\right) \in \underset{g\in\{1,\ldots,G\}}{\arg\min}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-p^{K}\left(z_{it}\right)'\beta^{K}-\alpha_{gt}\right)^{2}.$$

Lemma 7. *Let $\overline{M} := \max\{M^{-\frac{1}{2}}, M^{*}\}$ and $\eta^{*} := \frac{c^{2}}{16\left(\left(\sqrt{8}+2+\overline{M}\right)C+2\overline{M}\right)^{2}}$, where*

1. *M comes from Assumption 3.*
2. *c comes from Assumption 6.*
3. *C comes from Lemma 3.*
4. *M^{*} comes from Assumption 8.*

Under the Assumption 1-6 and 8,

$$\sup_{i\in\{1,\ldots,N\}}\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}=o_p\left(NT^{-\delta}\right),$$

as $N,T,K\to\infty$, where the ϕ function is defined in the Corollary 1's proof.

Proof. Let $D_{ig}:=\left\{\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-p_K\left(z_{it}\right)'\beta^K-\alpha_{\phi(g)t}\right)^2\leqslant\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-p_K\left(z_{it}\right)'\beta^K-\alpha_{\phi\left(g_i^0\right)t}\right)^2\right\}$.
Applying Lemma 3, then on the set $B\left(\eta^*,K,T\right)$ and $D_{ig}=1$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}&\leqslant-\frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2+C\sqrt{\eta}\left(\sqrt{8}+1\right)+C\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}\\&\quad+C\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}|x_{it}|^2\right)^{\frac{1}{2}}+2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}\\&\leqslant-\frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2+C\sqrt{\eta}\left(\sqrt{8}+1\right)+C\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}\\&\quad+C\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}|x_{it}|\right)+2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}\end{aligned}$$

Hence, on the set $B\left(\eta^*,K,T\right)$,

$$\begin{aligned}D_{ig}&\leqslant\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\leqslant-\frac{c}{3}+\sqrt{\eta}\left(C\left(\sqrt{8}+2\right)+2\overline{M}+C\overline{M}\right)\right\}+\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2>\frac{2}{3}c\right\}\\&\quad+\left\{\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}>\sqrt{\eta}\right\}+\left\{\frac{1}{T}\sum_{t=1}^{T}|x_{it}|>\overline{M}\right\}+\left\{\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}>\overline{M}\right\}\end{aligned}$$

Since $-\frac{c}{2}+\sqrt{\eta}\left(C\left(\sqrt{8}+2\right)+2\overline{M}+C\overline{M}\right)=-\frac{c}{12}$,

$$\begin{aligned}\cdot\left[D_{ig}\right]&\leqslant\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\leqslant-\frac{c}{12}\right)+\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2>\frac{2}{3}c\right)\\&\quad+\cdot\left(\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}>\sqrt{\eta}\right)+\cdot\left(\frac{1}{T}\sum_{t=1}^{T}|x_{it}|>\overline{M}\right)+\cdot\left(\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}>\overline{M}\right)\\&\leqslant\sup_{g\in\{1,\ldots,G^0\}\setminus\{g_i^0\};i\in\{1,\ldots,N\}}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\leqslant-\frac{c}{12}\right)\\&\quad+\sup_{i\in\{1,\ldots,N\};g\in\{1,\ldots,G^0\}\setminus\{g_i^0\}}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2>\frac{2}{3}c\right)\\&\quad+\cdot\left(\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}>\sqrt{\eta}\right)+\sup_{i\in\{1,\ldots,N\}}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}|x_{it}|>\overline{M}\right)+\sup_{i\in\{1,\ldots,N\}}\cdot\left(\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}>\overline{M}\right)\end{aligned}$$

1. For Assumption 1.3, $\cdot\left(\left|m-\left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}>\sqrt{\eta}\right)=0$ for large enough K.

2. From Assumption 8.4, $\sup_{i\in\{1,\ldots,N\}}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}|x_{it}|>\overline{M}\right)=o(T^{-\delta})$.

3. From Assumption 8.3 and 8.2, the $\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}$ process satisfies Lemma 4's conditions by taking $b_1=r_3$, $b_2=r_4$, $b_3=r_2$, and $b_4=2r_1\sup_{\alpha\in\mathcal{A}}|\alpha|$. Then applying Lemma 4 gives

$$\sup_{g\in\{1,\ldots,G^0\}\setminus\{g_i^0\};i\in\{1,\ldots,N\}}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\geqslant\frac{c}{12}\right)\leqslant f\left(T,\frac{c}{12},r_3,r_4,r_2,2r_1\sup_{\alpha\in\mathcal{A}}|\alpha|\right)$$

and

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\leqslant-\frac{c}{12}\right)\leqslant\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\right|\geqslant\frac{c}{12}\right).$$

Therefore, $\sup_{g\in\{1,\ldots,G^0\}\setminus\{g_i^0\};i\in\{1,\ldots,N\}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)\epsilon_{it}\leqslant-\frac{c}{12}\right)=o\left(T^{-\delta}\right)$.

4. From Assumption 8.1 and 8.3, the $\epsilon_{it}^2-\left[\epsilon_{it}^2\right]$ process satisfies Lemma 4's conditions by taking $b_1=r_3$, $b_2=r_4$, $b_3=r_1^{\frac{1}{2}}$, and $b_4=r_2^2$. Then applying Lemma 4 gives

$$\sup_{i\in\{1,\ldots,N\}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it}^2-\left[\epsilon_{it}^2\right]\right)\right|>\overline{M}^2-M^{-\frac{1}{2}}\right)\leqslant f\left(T,\overline{M}^2-M^{-\frac{1}{2}},r_3,r_4,r_1^{\frac{1}{2}},r_2^2\right)$$

and

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2>\overline{M}^2+\frac{1}{T}\sum_{t=1}^{T}\left(\left[\epsilon_{it}^2\right]-M^{-\frac{1}{2}}\right)\right)\leqslant\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it}^2-\left[\epsilon_{it}^2\right]\right)\right|>\overline{M}^2-M^{-\frac{1}{2}}\right).$$

Applying Jensen inequality with Assumption 3.3 and 3.4 yields $\left[\epsilon_{it}^2\right]-M^{-\frac{1}{2}}\leqslant 0$.

Therefore, $\sup_{i\in\{1,\ldots,N\}}\mathbb{P}\left(\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2\right)^{\frac{1}{2}}>\overline{M}\right)=o\left(T^{-\delta}\right)$.

5. Assume $g\neq g_i^0$. From Assumption 6.1, there exists a T^* such that $\frac{1}{T}\sum_{t=1}^{T}\left[\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2\right]>\frac{1}{3}c$, when $T>T^*$.

From Assumption 2 and 8.3, the $\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2-\left[\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2\right]$ process satisfies Lemma 4's conditions by taking $b_1=r_3$, $b_2=r_4$, $b_3=8\sup_{\alpha\in\mathcal{A}}|\alpha|$, and $b_4=1$. Then Lemma 4 gives

$$\begin{aligned}&\sup_{i\in\{1,\ldots,N\};g,g_i^0\in\{1,\ldots,G^0\}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2-\left[\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2\right]\right)\right|>\frac{1}{6}c\right)\\&\leqslant f\left(T,\frac{1}{6}c,r_3,r_4,8\sup_{\alpha\in\mathcal{A}}|\alpha|,1\right)\end{aligned}$$

and, for $T>T^*$,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2>\frac{2}{3}c\right)\leqslant\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2-\left[\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2\right]\right)\right|>\frac{1}{6}c\right).$$

Therefore, $\sup_{i\in\{1,\ldots,N\};g\in\{1,\ldots,G^0\}\setminus\{g_i^0\}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\phi(g)t}-\alpha^0_{\phi\left(g_i^0\right)t}\right)^2>\frac{2}{3}c\right)=o\left(T^{-\delta}\right)$.

Applying all the above yields $\sup_{i\in\{1,\ldots,N\};g\in\{1,\ldots,G^0\}\setminus\{g_i^0\}}\left[D_{ig}\right]=o_p\left(T^{-\delta}\right)$ on the set $B\left(\eta^*,K,T\right)$.
Since,

$$\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}\leqslant\sum_{g=1}^{G^0}\{g\neq g_i^0\}D_{ig}$$

then Markov inequality implies $\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}=o_p\left(T^{-\delta}\right)$.

$$\begin{aligned}\sup_{i\in\{1,\ldots,N\}}\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}&\leqslant\sum_{i=1}^{N}\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}\\&\leqslant No_p\left(T^{-\delta}\right)\\&=o_p\left(NT^{-\delta}\right)\end{aligned}$$

The second inequality follows from the above. □

Proof of Theorem 3. First of all, Lemma 5 and Markov inequality implies

$$\cdot\left(\sup_{i\in\{1,\ldots,N\}}\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}>0\right)=o(NT^{-\delta})\ (**).$$

Let $g\in\{1,\ldots,G\}$. Since $G=G^0$, the function ϕ has inverse by Corollary 1. Let $g^0:=\phi^{-1}(g)\in\{1,\ldots,G^0\}$.
Define $\hat{g}_i:=\hat{g}_i\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha}\right)$.

$$\begin{aligned}\cdot\left(\sup_{i\in\{1,\ldots,N\}}\left\{\hat{g}_i\neq\phi\left(g_i^0\right)\right\}>0\right)&\leqslant\cdot\left(\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha}\right)\notin B\left(\eta^*,K,T\right)\right)\\&\quad+\cdot\left(\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha}\right)\in B\left(\eta^*,K,T\right),\sup_{i\in\{1,\ldots,N\}}\left\{\hat{g}_i\neq\phi\left(g_i^0\right)\right\}>0\right)\\&=o(1)+o(NT^{-\delta})\text{ as }N,T,K\to\infty\end{aligned}$$

The last line comes from (**), Theorem 1's proof, and Corollary 1.
So therefore, $\cdot\left(\sup_{i\in\{1,\ldots,N\}}\left\{\hat{g}_i\neq\phi\left(g_i^0\right)\right\}>0\right)=o(1)+o(NT^{-\delta})=o(1)$, from Assumption 8.4. Then,

$$\begin{aligned}\cdot\left(H_{g^0}\subset H_g\right)&\geqslant 1-\cdot\left(\exists j\in H_{g^0},\{\hat{g}_i\neq g\}>0\right)\\&\leqslant 1-\cdot\left(\sup_{i\in\{1,\ldots,N\}}\left\{\hat{g}_i\neq\phi\left(g_i^0\right)\right\}>0\right)\\&=o(1)\end{aligned}$$

as $N,T,K\to\infty$.
If there exists a $j\in H_g\cap H_{g^0}^c$ then $\phi\left(g_j^0\right)\neq g$ because ϕ is injective from Corollary 1. Hence, $j\in H_g\cap H_{g^0}^c$ implies $\{\hat{g}_j\neq\phi\left(g_j^0\right)\}=1$. So,

$$\cdot\left(\exists j\in H_g\cap H_{g^0}^c\right)\leqslant\cdot\left(\sup_{i\in\{1,\ldots,N\}}\left\{\hat{g}_i\neq\phi\left(g_i^0\right)\right\}>0\right)=o(1).$$

In conclusion, $\cdot\left(H_g=H_{g^0}\right)=\cdot\left(H_{g^0}\subset H_g\right)+\cdot\left(H_g\cap H_{g^0}^c=\varnothing\right)=o(1)$. □

1. From here onward, the proof normalizes p^K by scaling it with $\left(\lim_{N\to\infty}\frac{\sum_{g=1}^{G^0}\psi_g^{pp,K}N_g}{N}\right)^{-\frac{1}{2}}$. This normalization is convenient in proving Theorem 4. The normalized basis will span the same subspace as p^K. So it will produce the same estimate of $\hat{m}$ and, hence, all previous results hold but with a different $\beta^{0,K}, \mu_K$ and ξ_K. To avoid introducing more notation, the new basis will still be referred to as p^K.

   After normalization $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left[\left(p^K(z_{it})-\psi_{g_i^0}^{p,K}(\alpha,1)\right)\left(p^K(z_{it})-\psi_{g_i^0}^{p,K}(\alpha,1)\right)'\right]=I_K$, where I_K is a $K\times K$ identity matrix. Assumption 9.2 (c) implies the new ξ_K and μ_K is just scaled by a constant. Thus satisfying rates on the original basis carries over to the new basis. Again, the same ξ_K and μ_K notations are used for the new basis. This change of basis is also in proofs of Newey (1997), Qi (2000), and Lee and Robinson (2016).

2. Define the concentrated criterion

$$\hat{Q}_c\left(\theta,\beta^K,\alpha\right):=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-p^K(z_{it})'\beta^K-\alpha_{\hat{g}_i(\theta,\beta^K,\alpha)t}\right)^2$$

   and its auxiliary criterion

$$\tilde{Q}_c\left(\theta,\beta^K,\alpha\right):=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-p^K(z_{it})'\beta^K-\alpha_{g_i^0t}\right)^2.$$

   Then the Oracle estimator solves the auxiliary criterion,

$$\left(\tilde{\theta},\tilde{\beta}^K,\tilde{\alpha}\right)\in\underset{\theta\in\Theta,\beta^K\in\mathcal{B}^K,\alpha\in\mathcal{A}^{G^0\times T}}{\arg\min}\tilde{Q}_c\left(\theta,\beta^K,\alpha\right).$$

3. After proving the uniform convergence of the two above criteria, I show $\sqrt{NT}\left(\hat{\theta}-\tilde{\theta}\right)=o_p(1)$. This step is completed by **Lemma 9**.

4. Let $\left(\check{\theta},\check{\beta}^K\right)\in\underset{\theta\in\Theta,\beta^K\in\mathcal{B}^K}{\arg\min}\ \bar{Q}_c\left(\theta,\beta^K\right)$, where

$$\bar{Q}_c\left(\theta,\beta^K\right):=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\left(x_{it}-\psi_{g_i^0}^{x}(\alpha,t)\right)'\left(\theta^0-\theta\right)-\left(p^K(z_{it})-\psi_{g_i^0}^{p^K}(\alpha,t)\right)'\beta^K+m(z_{it})-\psi_{g_i^0}^{m}(\alpha,t)+\epsilon_{it}\right)^2.$$

   I first show $\sqrt{NT}\left(\check{\theta}-\theta^0\right)$ is asymptotically normal.

5. Then I show $\sqrt{NT}\left(\check{\theta}-\tilde{\theta}\right)=o_p(1)$.

6. Theorem 4's proof is completed by combining step 3, step 4, and step 5.

7. **WLOG, the proof treats all estimators as the interior solution.** From the consistency of $\hat{\theta}$ and $\hat{\beta}^K$, Assumption 9.1 implies they are interior with asymptotic probability 1. Hence, it is sufficient to prove Theorem 4 under the assumption of $\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha}\right)$ being the interior solution. Then **Lemma 9**'s equivalence result implies $\left(\tilde{\theta},\tilde{\beta}^K,\tilde{\alpha}\right)$ can also be treated as interior for sake of proving Theorem 4. Also, $\check{\beta}$ and $\check{\theta}$ are obviously consistent. Thus after **Lemma 9**, all relevant estimators are treated as the interior solution without further mention.

Lemma 8. *Under Assumption 1-6 and 8,*

$$\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}=o_p\left(T^{-\delta}\right).$$

Proof. Lemma 5's proof shows $\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left\{\hat{g}_i\left(\theta,\beta^K,\alpha\right)\neq\phi\left(g_i^0\right)\right\}\leqslant\sum_{g=1}^{G^0}\{g\neq g_i^0\}D_{ig}$ and $\max_{g\neq g_i^0}\left[D_{ig}\right]=o_p\left(T^{-\delta}\right)$ uniformly across i. Then applying Markov inequality delivers the conclusion. □

Lemma 9. *Under Assumption 1-6 and 8,*

$$\sup_{\left(\theta,\beta^K,\alpha\right)\in B\left(\eta^*,K,T\right)}\left|\bar{Q}_c\left(\theta,\beta^K,\alpha\right)-\bar{Q}_c\left(\theta,\beta^K,\alpha\right)\right|=o_p\left(T^{-\delta}\right),$$

as $N,T,K\to\infty$.

Proof. Again, from Assumption 2, let C be large enough such that $2|\alpha| \leqslant C$ and $2\|\theta\| \leqslant C$, for $\theta \in \Theta$ and $\alpha \in \mathcal{A}$.

$$\begin{aligned}
\frac{1}{2}\left(\bar{Q}_c\left(\theta, \beta^K, \alpha\right) - \bar{Q}_c\left(\theta, \beta^K, \alpha\right)\right) = & \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\left(x_{it}'\left(\theta^0 - \theta\right)\right)\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}}_{A_1} \\
& + \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\left(p^K\left(z_{it}\right)'\left(\beta^{0,K} - \beta^K\right)\right)\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}}_{A_2} \\
& + \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\left(g\left(z_{it}\right) - p^K\left(z_{it}\right)'\beta^{0,K}\right)\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}}_{A_3} \\
& + \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\epsilon_{it}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}}_{A_4} \\
& + \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\left(\alpha_{g_i^0 t}^0 - \frac{\alpha_{g_i^0 t} + \alpha_{\hat{g}_i t}}{2}\right)\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}}_{A_5}.
\end{aligned}$$

1.

$$\begin{aligned}
|A_1| &\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\left(\alpha_{g_i^0 t} - \alpha_{\hat{g}_i t}\right)\right|\left|\left(x_{it}'\right)\right|\left|\left(\theta^0 - \theta\right)\right|\left\{\hat{g}_i \neq \phi\left(g_i^0\right)\right\} \\
&\leqslant M^{\frac{1}{4}} C^2 \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}.
\end{aligned}$$

The last line of inequality follows from Assumption 3.2.

2. Using a similar approach as (1),

$$|A_2| \leqslant C\sqrt{\eta}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}$$

on the set $B\left(\eta^*, K, T\right)$.

3. Using a similar approach as (1),

$$|A_3| \leqslant C\left|g - \left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}.$$

4. Using a similar approach as (1).

$$|A_4| \leqslant CM^{\frac{1}{4}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}.$$

The last line of inequality follows from Assumption 3.4.

5. Using a similar approach as (1),

$$|A_5| \leqslant \frac{1}{2}C^2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hat{g}_i\left(\theta, \beta^K, \alpha\right) \neq \phi\left(g_i^0\right)\right\}.$$

Hence, Markov inequality with Lemma 6 implies $A_j = o_p\left(T^{-\delta}\right)$ on the set $B\left(\eta^*, K, T\right)$. □

Lemma 10. *Under Assumption 1-5 and 6,* $\cdot\left(\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \in B\left(\eta^*, K, T\right)\right) \to 0$, *as* $N, T, K \to \infty$.

Proof. The proof is repeating Theorem 1's proof.
Define $\tilde{Q}_{c,p} := \tilde{Q}\left(\theta, \beta^K, \alpha, \gamma^0\right)$. Then repeating Lemma 1's argument shows

$$\sup_{\theta\in\Theta, \beta^K\in\mathcal{B}^K, \alpha\in\mathcal{A}^{G^0\times T}} \left|\tilde{Q}_c - \tilde{Q}_{c,p}\right| = o_p(1),$$

as $N, T, K \to \infty$.
With this uniform convergence, adapting the argument of Theorem 1's Step 1 can lead to $\left|\tilde{\theta} - \theta^0\right| = o_p(1)$, $\left|\tilde{\beta}^K - \beta^{0,K}\right| = o_p(1)$, and $\xi_K\left|\beta^{0,K} - \tilde{\beta}^K\right| = o_p(1)$. Then adapting the argument of Theorem 1's Step 2 and Corollary 1 can lead to $\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\alpha}_{g^0 t} - \alpha_{g^0 t}\right)^2 = o_p(1)$, for any $g^0 \in \{1, ..., G^0\}$.
In conclusion, $\cdot\left(\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \in B\left(\eta^*, K, T\right)\right) \leqslant \cdot\left(\left|\tilde{\theta} - \theta^0\right|^2 \leqslant \eta\right) + \cdot\left(\left|\tilde{\beta}^K - \beta^{0,K}\right|^2 \leqslant \eta\right)$
$+ \sum_{g^0=1}^{G^0} \cdot\left(\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{\alpha}_{g^0 t} - \alpha^0_{g^0 t}\right)^2 \leqslant \eta\right) + \cdot\left(\xi_K^2\left|\tilde{\beta}^K - \beta^{0,K}\right|^2 \leqslant \eta\right) = o_p(1)$. □

Lemma 11. *Under Assumption 1-6 and 8, as* $N, T, K \to \infty$,

1. $\sqrt{NT}\left(\hat{\theta} - \tilde{\theta}\right) = o_p(1)$,
2. $\sqrt{NT}\left(\hat{\beta}^K - \tilde{\beta}^K\right) = o_p(1)$, *and*
3. $\sum_{t=1}^{T}\left(\tilde{\alpha}_{\phi\left(g^0\right)t} - \tilde{\alpha}_{g^0 t}\right)^2 = o_p\left(T^{1-\delta}\right)$, *for all* $g^0 \in \{1, ..., G^0\}$ *and* $t \in \{1, ..., T\}$.

Proof. Combining Theorem 1, Lemma 7, and 8 yield,

$$\hat{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) = o_p\left(T^{-\delta}\right) \text{ and } \hat{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) - \tilde{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) = o_p\left(T^{-\delta}\right),$$

as $\cdot\left(\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \in B\left(\eta^*, K, T\right)\right) \to 0$ and $\cdot\left(\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) \in B\left(\eta^*, K, T\right)\right) \to 0$.
Hence, as $N, T, K \to \infty$,

$$\begin{aligned}
\tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \tilde{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) &= \hat{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) + \tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \hat{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \\
&\quad + \hat{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) - \tilde{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \\
&= o_p\left(T^{-\delta}\right) + \tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \hat{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) + \hat{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \hat{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) \\
&= o_p\left(T^{-\delta}\right)
\end{aligned}$$

The last line follows from the fact $\tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) \geqslant \tilde{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right)$ but $\hat{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) \leqslant \hat{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right)$.
With $\tilde{Q}_c\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}\right) - \tilde{Q}_c\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}\right) = o_p\left(T^{-\delta}\right)$, a similar argument to Step 1 of Theorem 1's proof leads to $\hat{\theta} = \tilde{\theta} + o_p\left(T^{-\frac{\delta}{2}}\right)$ and $\hat{\beta}^K = \tilde{\beta}^K + o_p\left(T^{-\frac{\delta}{2}}\right)$, as $K, N, T \to \infty$. As $\frac{N}{T^{\delta-1}} \to 0$, $\sqrt{NT}\left(\hat{\theta} - \tilde{\theta}\right) = o_p(1)$, and $\sqrt{NT}\left(\hat{\beta}^K - \tilde{\beta}^K\right) = o_p(1)$,
And a similar argument to Theorem 1's Step 2 leads to $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right)^2 = o_p\left(T^{-\delta}\right)$. Then a similar argument to Corollary 1's proof leads to $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right)^2 = o_p\left(T^{-\delta}\right)$. In conclusion, $\sum_{t=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right)^2 = o_p\left(T^{1-\delta}\right)$. □

The following notations help to prove the Theorem 4's asymptotic normality.

Let $Y := \begin{bmatrix} y_{g_1^0 1} \\ \vdots \\ y_{g_N^0 T} \end{bmatrix}$, $Y^{\psi} := \begin{bmatrix} \left(\psi^x_{g_1^0}(\alpha,1)\right)'\theta + \psi^m_{g_1^0}(\alpha,1) \\ \vdots \\ \left(\psi^x_{g_N^0}(\alpha,T)\right)'\theta + \psi^m_{g_N^0}(\alpha,T) \end{bmatrix}$, $\overline{Y} := Y - Y^{\psi}$, $X := \begin{bmatrix} x'_{11} \\ \vdots \\ x'_{NT} \end{bmatrix}$, $X^{\psi} := \begin{bmatrix} \left(\psi^x_{g_1^0}(\alpha,1)\right)' \\ \vdots \\ \left(\psi^x_{g_N^0}(\alpha,T)\right)' \end{bmatrix}$, $\overline{X} := X - X^{\psi}$,

$\mathcal{M} := \begin{bmatrix} m(z_{11}) \\ \vdots \\ m(z_{NT}) \end{bmatrix}$, $\mathcal{M}^{\psi} := \begin{bmatrix} \psi^m_{g_1^0}(\alpha,1) \\ \vdots \\ \psi^m_{g_N^0}(\alpha,T) \end{bmatrix}$, $\overline{\mathcal{M}} := \mathcal{M} - \mathcal{M}^{\psi}$, $\mathcal{E} := \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{NT} \end{bmatrix}$, $P := \begin{bmatrix} p^K(z_{11})' \\ \vdots \\ p^K(z_{NT})' \end{bmatrix}$, $P^{\psi} := \begin{bmatrix} \left(\psi^{p,K}_{g_1^0}(\alpha,1)\right)' \\ \vdots \\ \left(\psi^{p,K}_{g_N^0}(\alpha,T)\right)' \end{bmatrix}$, $\overline{P} := P - P^{\psi}$,

$\mathbf{M} := I_{NT} - \overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'$, $\zeta := \left(\theta'\left(\beta^K\right)'\right)'$, $\overline{x_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} x_{jt}$, $\overline{p^K_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} p^K(z_{jt})$, $\overline{m_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} m(z_{jt})$,

$\overline{\epsilon_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} \epsilon_{jt}$, $\overline{y_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} y_{jt}$, $\overline{\sigma_{gt}} := \frac{1}{N_g}\sum_{j:g_j^0=g} \sigma(x_{it}, z_{jt})$, $\mathbf{p}^K(z_{it}) := p^K(z_{it}) - \psi^{p,K}_{g_i^0}(\alpha,t)$, $\mathbf{x}_{it} := x_{it} - \psi^x_{g_i^0}(\alpha,t)$,

$\mathbf{m}(z_{it}) := m(z_{it}) - \psi^m_{g_i^0}(\alpha,t)$, $tr(\cdot)$ is the trace function, and $\cdot_Z\left[\overline{X}\right] := \begin{bmatrix} \cdot[\mathbf{x}_{11}|z_{11}]' \\ \vdots \\ \cdot[\mathbf{x}_{NT}|z_{NT}]' \end{bmatrix}$.

Furthermore, define the relation $A \lesssim B$ to imply there exists a a constant C such that $A \leqslant CB$.

Lemma 12. *Let* $\mathbf{v}^{g_i}(z_{it}) := \cdot[\mathbf{x}_{it} \mid z_{it}]$. *Under Assumption 1, Assumption 6, and Assumption 9,* $\left|\mathbf{v}^g_j - \left(\mathbf{p}^K\right)'\beta^K_{x,j}\right|_{\infty,\mathcal{Z}} = O\left(K^{-\mu} + \sqrt{K}\Pi^x N^{-1}\right)$ *and* $\left|\mathbf{m} - \left(\mathbf{p}^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}} = O\left(K^{-\mu} + \sqrt{K}\Pi_K N^{-1}\right)$.

Proof. The lemma states, the group-wise demeaned model inherit the rate convergence of series approximation of the conditional mean and m function.

$$\begin{aligned}
\left|\mathbf{v}_j - \left(\mathbf{p}^K\right)'\beta^K_{x,j}\right|_{\infty,\mathcal{Z}} &\leqslant \left|v_j - c_{vj} - \left(p^K\right)'\beta^K_{x,j}\right|_{\infty,\mathcal{Z}} + \sum_{g=1}^{G^0}\sup_{1\leqslant t\leqslant T}\left|\cdot\left[\left[\psi^x_g(\alpha,t)\right]_j - \frac{\sum_{i:g_i^0=g}\cdot\left[x_{it,j}|\alpha\right]}{N_g}|z\right]\right| \\
&+ \sum_{g=1}^{G^0}\sup_{1\leqslant t\leqslant T}\left|\cdot\left[\psi^{p,K}_g(\alpha,t) - \frac{\sum_{i:g_i^0=g}\cdot\left[p^K(z_{it})|\alpha\right]}{N_g}|z\right]\right|\left|\beta^K_{x,j}\right| \\
&+ \sum_{g=1}^{G^0}\sup_{1\leqslant t\leqslant T}\left|\cdot\left[\frac{\sum_{i:g_i^0=g}\cdot\left[x_{it,j} - c_{vj} - p^K(z_{it})'\beta^K_{xj}|\alpha\right]}{N_g}|z\right]\right| \\
&\leqslant O\left(K^{-\mu}\right) + 2\left(1+\sqrt{K}\Pi^x\right)G^0\frac{C^{xp}}{N_g} + \sum_{g=1}^{G^0}\sup_{1\leqslant t\leqslant T}\left|\cdot\left[\frac{\sum_{i:g_i^0=g}\cdot\left[x_{it,j} - c_{vj} - p^K(z_{it})'\beta^K_{xj}|\alpha\right]}{N_g}|z\right]\right|
\end{aligned}$$

The last line uses the rate provided in Assumption 1 and Assumption 9.

$$\begin{aligned}
\sum_{g=1}^{G^0}\sup_{1\leqslant t\leqslant T}\left|\cdot\left[\frac{\sum_{i:g_i^0=g}\cdot\left[x_{i1,j} - c_{vj} - p^K(z_{i1})'\beta^K_{xj}|\alpha\right]}{N_g}|z\right]\right| &\leqslant G^0\sup_{i\in\{1,\ldots,N\}}\left|v_j(\alpha) - c_{vj} - \left(p^K\right)'\beta^K_{x,j}\right|_{\infty,\mathcal{Z}} \\
&\leqslant MG^0O(K^{-\mu})
\end{aligned}$$

The last line follows from Cauchy-Schwarz inequality.
A similar argument can provided for the $\left|\mathbf{m}-\left(\mathbf{p}^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}}$ case. □

Lemma 13. *Under Assumption 1, 5, 6, and 9,* $\frac{\widetilde{P}'\widetilde{P}}{NT} = I_K + o_p(1)$, *as* $N, T, K \to \infty$.

Proof. This is Newey (1997)'s Theorem 1 with relaxing the independence and identically distributed assumption.
In summation form,

$$\frac{\widetilde{P}'\widetilde{P}}{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right)$$

Let $\cdot := \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\,|\alpha\right]\right)^{-\frac{1}{2}}$ and $\mathbf{p}^K(z_{it}) := \cdot\mathbf{p}^K(z_{it})$.
Let I_{ls} denote the lsth entry in the matrix I_K and $\psi^{p,K}_{g_i^0,l}$ be the lth entry of $\psi^{p,K}_{g_i^0}$.

$$\begin{aligned}
&\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right)-I_K\right|^2|\alpha\right] = \cdot\left[\sum_{l=1}^{K}\sum_{s=1}^{K}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\right)-I_{ls}\right)^2|\alpha\right]\\
&=\sum_{l=1}^{K}\sum_{s=1}^{K}\frac{\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\Sigma_{t'=1}^{T}Cov\left(\mathbf{p}^K_{\cdot,l}\left(z_{it'}\right)\mathbf{p}^K_{\cdot,s}\left(z_{it'}\right),\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\,|\alpha\right)}{N^2T^2}\\
&+\sum_{l=1}^{K}\sum_{s=1}^{K}\frac{\Sigma_{i=1}^{N}\Sigma_{j\neq i}^{N}\Sigma_{t=1}^{T}\Sigma_{t'=1}^{T}Cov\left(\mathbf{p}^K_{\cdot,l}\left(z_{it'}\right)\mathbf{p}^K_{\cdot,s}\left(z_{it'}\right),\mathbf{p}^K_{\cdot,l}\left(z_{jt}\right)\mathbf{p}^K_{\cdot,s}\left(z_{jt}\right)|\alpha\right)}{N^2T^2}
\end{aligned}$$

The second equality comes from $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\cdot\left[\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\,|\alpha\right] = I_{ls}$.
From conditional independence over i, j,

$$\sum_{l=1}^{K}\sum_{s=1}^{K}\frac{\Sigma_{i=1}^{N}\Sigma_{j\neq i}^{N}\Sigma_{t=1}^{T}\Sigma_{t'=1}^{T}Cov\left(\mathbf{p}^K_{\cdot,l}\left(z_{it'}\right)\mathbf{p}^K_{\cdot,s}\left(z_{it'}\right),\mathbf{p}^K_{\cdot,l}\left(z_{jt}\right)\mathbf{p}^K_{\cdot,s}\left(z_{jt}\right)|\alpha\right)}{N^2T^2} = 0.$$

And from the normalization by $\cdot$, whenever $l \neq s$,

$$\frac{\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}Cov\left(\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it}),\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\,|\alpha\right)}{N^2T^2} = 0.$$

From Davydov inequality and Assumption 9.1(b), whenever $l = s$,

$$\begin{aligned}
\left|Cov\left(\mathbf{p}^K_{\cdot,l}(z_{is})\,\mathbf{p}^K_{\cdot,s}(z_{is}),\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\,|\alpha\right)\right| &\lesssim 12e^{-\frac{1}{3}r_3t^{r_4}}\cdot\left[\left|\mathbf{p}^K_{\cdot,l}(z_{is})\,\mathbf{p}^K_{\cdot,s}(z_{is})\right|^3|\alpha\right]^{\frac{1}{3}}\cdot\left[\left|\mathbf{p}^K_{\cdot,l}(z_{it})\,\mathbf{p}^K_{\cdot,s}(z_{it})\right|^3|\alpha\right]^{\frac{1}{3}}\\
&\lesssim e^{-\frac{1}{3}r_3|t-s|^{r_4}}\xi_K^2,
\end{aligned}$$

because Assumption 9.3 provides $|\mathbf{p}^K_{\cdot,s}(z_{i1})| \lesssim \xi_K$.
With the iterated law of expectation,

$$\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right)-I_K\right|^2\right] \lesssim \left(\sum_{t=1}^{T}\sum_{s=1}^{T}e^{-\frac{1}{3}r_3|t-s|^{r_4}}\right)\frac{K\xi_K^2}{NT^2}.$$

Assumption 5.2 provides $\frac{\xi_K^6K^2\text{Π}_K^2}{NT} \to 0$. Assumption 1 sends $\xi_K \to \infty$ and sets Π_K as uniformly bounded away from 0. Therefore, $\frac{\xi_K^2K}{NT} \to 0$.
As $\sum_{s=1}^{T}e^{-\frac{1}{3}r_3|t-s|^{r_4}}$ is convergent by the Ratio Test,

$$\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right)-I_K\right|^2\right] = o_p(1).$$

Similarly,

$$\begin{aligned}
&\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\,|\alpha\right]-\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right]\right|\right]\\
&\leqslant\frac{1}{N}\sum_{i=1}^{N}\cdot\left[\left|\frac{1}{T}\sum_{t=1}^{T}\left[\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\,|\alpha\right]-\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right]\right]\right|\right]\\
&\lesssim\frac{K\xi_K}{T}
\end{aligned}$$

when the last line follows a Davydov inequality argument and Assumption 9.1, like above. Recall the law of total covariance. As the Davydov bounds for both conditional and unconditional, it also applies to the above.

From the starting normalization and Assumption 9, $\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right]-I_K\right|=o(1)$. The last line follows from the normalization and Assumption 9. Then by triangle inequality, $\left|\cdot^{-1}-I_K\right|=o_p(1)$, as $\frac{K\xi_K}{T}\to 0$. Hence, by continuity, $\left|\cdot-I_K\right|=o_p(1)$.

In conclusion,

$$\begin{aligned}
&\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\,|\alpha\right]-I_K\right]\right|\\
&\leqslant\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\,|\alpha\right]-\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right]\right]\right|+\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[I_K-\cdot\left[\mathbf{p}^K(z_{it})\,\mathbf{p}^K(z_{it})'\right]\right]\right|\\
&=o_p(1).
\end{aligned}$$

□

Lemma 14. *Define* $\mathbf{b}^K:=\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\overline{\mathcal{M}}$, $\mathbf{b}_x^K:=\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'_{\cdot Z}\left[\overline{X}\right]$, *and* $\beta_x^K(\alpha):=\left(\beta_{x,1}^K(\alpha),\ldots,\beta_{x,d_2}^K(\alpha)\right)$. *Under Assumption 1, 5, 6, and 9,* $\left|\beta_x^K-\mathbf{b}_x^K\right|=O_p\left(K^{-\mu}+\sqrt{K}\Pi^x{}_N{}^{-1}\right)$ *and* $\left|\beta^{0,K}-\mathbf{b}^K\right|=O_p\left(K^{-\mu}+\sqrt{K}\Pi_K{}_N{}^{-1}\right)$.

Proof. The argument is similar to Qi (2000)'s Lemma A.2's proof.
Let $\overline{X}_j$ be the jth column of matrix $\overline{X}$.

$$\begin{aligned}
\left|\beta_x^K-\mathbf{b}_x^K\right|^2&\leqslant\sum_{j=1}^{d_2}\left|\beta_{x,j}^K-\mathbf{b}_{x,j}^K\right|^2\\
&\leqslant\sum_{j=1}^{d_2}tr\left(\left(\left[\begin{array}{c}\mathbf{v}^{\mathbf{g}}{}_j(z_{11})\\ \vdots\\ \mathbf{v}^{\mathbf{g}}{}_j(z_{NT})\end{array}\right]-\overline{P}\beta_{x,j}^K\right)'\overline{P}\left(\frac{1}{NT}\overline{P}'\overline{P}\right)^{-1}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\left(\left[\begin{array}{c}\mathbf{v}^{\mathbf{g}}{}_j(z_{11})\\ \vdots\\ \mathbf{v}^{\mathbf{g}}{}_j(z_{NT})\end{array}\right]-\overline{P}\beta_{x,j}^K\right)\right)\\
&\leqslant O_p(1)\sum_{j=1}^{d_2}tr\left(\frac{1}{NT}\left(\left[\begin{array}{c}\mathbf{v}^{\mathbf{g}}{}_j(z_{11})\\ \vdots\\ \mathbf{v}^{\mathbf{g}}{}_j(z_{NT})\end{array}\right]-\overline{P}\beta_{x,j}^K\right)'\left(\left[\begin{array}{c}\mathbf{v}^{\mathbf{g}}{}_j(z_{11})\\ \vdots\\ \mathbf{v}^{\mathbf{g}}{}_j(z_{NT})\end{array}\right]-\overline{P}\beta_{x,j}^K\right)\right)\\
&\leqslant O_p(1)\,d_2\sup_{g=1,\ldots,G^0;j=1,\ldots,d_2}\left|\mathbf{v}^{\mathbf{g}}{}_j-\left(\mathbf{p}^K\right)'\beta_{x,j}^K\right|_{\infty,\mathcal{Z}}^2\\
&\leqslant d_2M\left(O_p\left(K^{-\mu}+\sqrt{K}\Pi^x{}_N{}^{-1}\right)\right)^2
\end{aligned}$$

The third line comes from using **Lemma 11** and $\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'$ being idempotent and having rank K. And the last line's bound uses Assumption 3.
A similar argument applies to $\left|\beta^{0,K}-\mathbf{b}^K\right|=O_p\left(K^{-\mu}+\sqrt{K}\Pi_K{}_N{}^{-1}\right)$. □

Lemma 15. *Under Assumption 9,* $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\cdot\left[x_{it}|z_{it}\right]-\psi(\alpha,t)+\cdot\left[\psi^x(\alpha,t)\,|z_{it}\right]\right)\epsilon_{it}\Rightarrow N\left(0,\psi^{x\epsilon}\right)$.

Proof. Define $w_{it}:=\left(x_{it}-\cdot\left[x_{it}|z_{it}\right]-\psi(\alpha,t)+\cdot\left[\psi(\alpha,t)\,|z_{it}\right]\right)\epsilon_{it}$ and $\Sigma_i^w:=\sum_{t=1}^{\infty}\cdot\left[w_{i1}w_{it}'\right]$.
The proof proceeds in the following steps. Let $v\in\cdot^{d_2}$ and $J_{i,N,T}:=\frac{\sum_{t=1}^{T}v'w_{it}}{\sqrt{NT}}$.

1. Show $\sum_{i=1}^{N}Var\left(J_{i,N,T}\mid\alpha\right)=v'\psi^{x\epsilon}v+o_p(1)$.

2. Show $N \sum_{i=1}^{N} \cdot \left[J_{i,N,T}^4 \mid \alpha\right] = O_p(1)$.

3. Apply the Lyapunov condition to show $\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}}{\sqrt{NT}} \Rightarrow N(0, \Sigma(\alpha))$, as $N, T \to \infty$, for some positive definite $\Sigma(\alpha)$.

4. Using $\psi^{x\epsilon} = \Sigma(\alpha) + o_p(1)$ to conclude.

On the (1) step:
By Davydov inequality, Cauchy-Schwarz inequality and $\sup_{t\in\{1,\ldots,T\}} \sup_{i\in\{1,\ldots,N\}} \cdot \left[|w_{it}|^5\right] \leqslant M^m$,

$$\left|Cov\left(v'w_{i1}, v'w_{i1+t}\right)\right| \lesssim (\rho_i(t))^{\frac{1}{3}} \left(M^m\right)^{\frac{2}{5}} \|v\|^2.$$

Hence, $\sup_{i\in\{1,\ldots,N\}} \left|Cov\left(v'w_{i1}, v'w_{i1+t}\right)\right| \lesssim \sup_{i\in\{1,\ldots,N\}} (\rho_i(t))^{\frac{1}{3}} \lesssim e^{-\frac{1}{3}r_3 t^{r_4}}$.
Via substitution and by expanding out,

$$\begin{aligned}\left|Var\left(J_{i,N,T}\right) - \frac{1}{N}v'\Sigma_i^w v\right| &= \frac{1}{N}\Big| Cov\left(v'w_{i1}, v'w_{i1}\right) + 2\sum_{t=1}^{T-1}\left(1-\frac{t}{T}\right)Cov\left(v'w_{i1}, v'w_{i1+t}\right) \\ &\quad - Cov\left(v'w_{i1}, v'w_{i1}\right) - 2\sum_{t=1}^{\infty} Cov\left(v'w_{i1}, v'w_{i1+t}\right)\Big| \\ &\leqslant 2\frac{1}{N}\left(\sum_{t=T}^{\infty}\left|Cov\left(v'w_{i1}, v'w_{i1+t}\right)\right| + \sum_{t=1}^{\infty}\left|Cov\left(v'w_{i1}, v'w_{i1+t}\right)\right|\frac{t}{T}\{t \leqslant T-1\}\right)\end{aligned}$$

The first line uses the stationary assumption. And the second comes from the triangle inequality.
Therefore, $\sup_{i\in\{1,\ldots,N\}}\left|Var\left(J_{i,N,T}\right) - \frac{1}{N}v'\Sigma_i^w v\right| \lesssim \frac{2}{N}\left(\sum_{t=T}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}} + \sum_{t=1}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}}\frac{t}{T}\{t \leqslant T-1\}\right)$. By the Ratio Test, $\sum_{t=1}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}}$ is convergent. Therefore, $\lim_{T\to\infty}\sum_{t=1}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}} = 0$ and, by Dominated Convergence with the dominating function $e^{-\frac{1}{3}r_3 t^{r_4}}$,

$$\lim_{T\to\infty}\sum_{t=1}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}}\frac{t}{T}\{t \leqslant T-1\} = \sum_{t=1}^{\infty} e^{-\frac{1}{3}r_3 t^{r_4}} \lim_{T\to\infty}\frac{t}{T}\{t \leqslant T-1\} = 0.$$

Thus the step deduces $\sum_{i=1}^{N} Var\left(J_{i,N,T}\right) = v'\psi^{x\epsilon}v + o(1)$, as $N, T \to \infty$. Then by Markov inequality and law of total variance, $\sum_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right) = v'\psi^{x\epsilon}v + o_p(1)$.
On the (2) step:
Recall $\sup_{i\in\{1,\ldots,N\}} \cdot \left[|w_{it}|^5\right] \leqslant M^m$ and $\sup_{i\in\{1,\ldots,N\}} \rho_i(t) \leqslant e^{-r_3 t^{r_4}} = O\left(t^{-10}\right)$[2].
With these conditions, the proof can invoke Fan and Q.Yao (2003)'s Proposition 2.7 to apply Doukhan and Louhichi (1999)'s Theorem 1. And Doukhan and Louhichi (1999)'s Theorem 1 provides a constant C^J such that $\sup_{i\in\{1,\ldots,N\}} \cdot \left[J_{i,N,T}^4\right] \leqslant \frac{C^J}{N^2}$.
Then $\sum_{i=1}^{N} \cdot \left[J_{i,N,T}^4\right] \leqslant \frac{C^J}{N}$. With Markov inequality and the iterated law of expectation, $N\sum_{i=1}^{N} \cdot \left[J_{i,N,T}^4 \mid \alpha^0\right] = O_p(1)$.
On the (3) step: Let $\cdot > 0$ and consider the event: $\cdot = \left\{N\sum_{i=1}^{N} \cdot \left[J_{i,N,T}^4 \mid \alpha\right] < \right\} \bigcap \left\{\frac{v'\psi^{x\epsilon}v}{2} < \sum_{i=1}^{N} Var\left(J_{i,N,T}|\alpha\right)\right\}$. On this event $\cdot$,

$$\frac{\sum_{i=1}^{N} \cdot \left[J_{i,N,T}^4 \mid \alpha\right]}{\left(\sum_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)\right)^2} \leqslant \frac{1}{N}\frac{4}{(v'\psi^{x\epsilon}v)^2} \to 0$$

as $N, T \to \infty$.
So the Lyapunov condition holds. From $J_{i,N,T}$'s conditional (on α) independence over i, $\frac{\sum_{i=1}^{N} J_{i,N,T}}{\sqrt{\sum_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)}} \Rightarrow N(0,1)$, as $N, T \to \infty$, on the event $\cdot$. With Step (1) and Step (2), the event $\cdot$ occurs with probability converging to 1 as $N, T \to \infty$, i.e. $\{\cdot\} = 1 + o_p(1)$ from Markov

[2] As $t \to \infty$, $t^{10}e^{-\frac{1}{5}r_3 t^{r_4}} \to 0$ because log is a slowly varying function.

inequality. Then applying the continuous mapping theorem on the decomposition

$$\frac{\Sigma_{i=1}^{N} J_{i,N,T}}{\sqrt{\Sigma_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)}} = \frac{\Sigma_{i=1}^{N} J_{i,N,T}}{\sqrt{\Sigma_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)}} \{.\} + (1 - \{.\}) \frac{\Sigma_{i=1}^{N} J_{i,N,T}}{\sqrt{\Sigma_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)}}$$

yields $\frac{\Sigma_{i=1}^{N} J_{i,N,T}}{\sqrt{\Sigma_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right)}} \Rightarrow N(0,1)$.

Then by the Cramer-Wold device, $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it} \Rightarrow N\left(0, \Sigma^{W}(\alpha)\right)$, for some positive definite $\Sigma^{W}(\alpha)$.

On the (5) step:

From Step (3)'s result, $\sum_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right) = \Sigma(\alpha) + o_p(1)$, and from Step (1)'s result, $\sum_{i=1}^{N} Var\left(J_{i,N,T} \mid \alpha\right) = \psi^{x\epsilon} + o_p(1)$. Then it follows that $\psi^{x\epsilon} = \Sigma(\alpha) + o_p(1)$. Then by continuous mapping theorem,

$$\left[\psi^{x\epsilon}\right]^{-1} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it} \Rightarrow N\left(0, I_{d_2}\right).$$

□

Lemma 16. *Under Assumption 1, 3, 5, 6, 8 and 9,* $\frac{1}{NT} \overline{X}' \mathbf{M} \overline{X} = \sum_{g=1}^{G^0} \kappa_g \psi_g^{xz} + o_p(1)$.

Proof.

$$\begin{aligned}
\mathbf{M}\overline{X} &= \overline{X} - \overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\overline{X} \\
&= \overline{X} - {}_{Z}\left[\overline{X}\right] + \left({}_{Z}\left[\overline{X}\right] - \overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'{}_{Z}\left[\overline{X}\right]\right) - \overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)
\end{aligned}$$

$$\begin{aligned}
\left[\left|\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)\right|^2 \mid Z, \alpha\right] &= tr\left(\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\left[\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)' \mid Z, \alpha\right]\right) \\
&\lesssim TO\left(M^m\right) tr\left(\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\right) \\
&= O(KT)
\end{aligned}$$

Note $\left[\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)' \mid Z, \alpha\right]$ is block diagonal with all its submatrices being $T \times T$ and entry bounded by M^m. Thus its largest eigenvalue is bounded above by $M^m T$. Since it is also symmetric, the spectral decomposition can replace it in the trace with its largest eigenvalue because $\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'$ is a positive definite matrix. Hence, the second line follows. The third line uses the fact of $\overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'$ as idempotent and having rank K, iterated law of expectation, and Assumption 9's bound on the conditional variance.

$$\begin{aligned}
\left|\left({}_{Z}\left[\overline{X}\right] - \overline{P}\left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'{}_{Z}\left[\overline{X}\right]\right)\right| &\leqslant \left|{}_{Z}\left[\overline{X}\right] - \overline{P}\beta_x^K\right| + \left|\overline{P}\left(\beta_x^K - \mathbf{b}_x^K\right)\right| \\
&\leqslant M\sqrt{NT}O\left(K^{-\mu} + \sqrt{K}\Pi^x N^{-1}\right) + \sqrt{NT}tr\left(\left(\beta_x^K - \mathbf{b}_x^K\right)'\frac{\overline{P}'\overline{P}}{NT}\left(\beta_x^K - \mathbf{b}_x^K\right)\right)^{\frac{1}{2}} \\
&\leqslant \sqrt{NT}O_p\left(K^{-\mu} + \sqrt{K}\Pi^x N^{-1}\right)
\end{aligned}$$

The second last line uses **Lemma 10**. The last line uses **Lemma 12**. Thus, $\frac{1}{\sqrt{NT}}\mathbf{M}\overline{X} = \overline{X} - {}_{Z}\left[\overline{X}\right] + O_p\left(\frac{\sqrt{K}}{\sqrt{N}}\right) + O_p\left(K^{-\mu} + \sqrt{K}\Pi^x N^{-1}\right)$. As $\overline{X} - {}_{Z}\left[\overline{X}\right] = O_p\left(\sqrt{NT}\right)$, therefore

$$\frac{1}{NT}\overline{X}'\mathbf{M}\overline{X} = \frac{1}{NT}\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right)'\left(\overline{X} - {}_{Z}\left[\overline{X}\right]\right) + o_p(1)$$

based on the rates in Assumption 5 and 9.

Define $\mathbf{xz}_{it} := \left(x_{it} - \cdot[x_{it} \mid z_{it}] - \psi(\alpha,t) + \cdot[\psi(\alpha,t) \mid z_{it}]\right)\left(x_{it} - \cdot[x_{it} \mid z_{it}] - \psi^{x}(\alpha,t) + \cdot\left[\psi^{x}(\alpha,t) \mid z_{it}\right]\right)'$.

$$\begin{aligned}\frac{1}{NT}\left(\widetilde{X} - \cdot_{Z}\left[\widetilde{X}\right]\right)'\left(\widetilde{X} - \cdot_{Z}\left[\widetilde{X}\right]\right) - \sum_{g=1}^{G^0}\kappa_g\psi_g^{xz} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{xz}_{it} - \cdot[\mathbf{xz}_{it}]\right)\\ &\quad + \sum_{g=1}^{G^0}\frac{N_g}{N}\left(\psi_g^{xz} - \frac{1}{N_g}\sum_{i:g_i^0=g}\cdot[\mathbf{xz}_{it}]\right) + \sum_{g=1}^{G^0}\left(\frac{N_g}{N} - \kappa_g\right)\psi_g^{x}.\end{aligned}$$

By Assumption 9, $\sum_{g=1}^{G}\frac{N_g}{N}\left(\psi_g^{xz} - \frac{1}{N_g}\sum_{i:g_i^0=g}\cdot[\mathbf{xz}_{it}]\right) = o(1)$. With Assumption 6, $\frac{N_g}{N} \to \kappa_g$, as $N \to \infty$.

$$\begin{aligned}\cdot\left[\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{xz}_{it} - \cdot[\mathbf{xz}_{it}]\right)\right|^2\right] &= tr\left(\cdot\left[\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\mathbf{xz}_{it} - \cdot[\mathbf{xz}_{it}]\right)\left(\mathbf{xz}_{it} - \cdot[\mathbf{xz}_{it}]\right)'\right]\right)\\ &= \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{t=1}^{T}tr\left(Cov\left(\mathbf{xz}_{i1}, \mathbf{xz}_{it}\right)\right)\\ &\lesssim \frac{1}{N^2T}\sum_{i=1}^{N}\sum_{t=1}^{T}d_2\rho_i(t)^{\frac{1}{3}}\cdot\left[\left|\mathbf{xz}_{i1}\right|^3\right]^{\frac{2}{3}}\\ &\lesssim \frac{1}{N}\sum_{t=1}^{T}\frac{1}{T}d_2e^{-r_3t^{r_4}}\left(\max\left\{1, M^m\right\}\right)^{\frac{2}{3}}\end{aligned}$$

The second line follows from the stationary assumption, the iterated law of expectation, and the conditional independence assumption. The third lines uses Davydov inequality and the stationary assumption. The fourth lines uses the tail bound in Assumption 8. By the Ratio Test, $\sum_{t=1}^{T}\frac{1}{T}d_2e^{-r_3t^{r_4}} = O(1)$. Hence the fourth line is $o(1)$. □

Lemma 17. *Under Assumption 1, 3, 5, 6, and 9,* $\sqrt{NT}\left(\breve{\theta} - \theta^0\right) \Rightarrow N\left(0, \Sigma_\theta\right)$.

Proof. By Frisch–Waugh–Lovell theorem, $\breve{\theta} - \theta^0 = \left(\widetilde{X}\mathbf{M}\widetilde{X}\right)^{-1}\widetilde{X}\mathbf{M}\widetilde{\mathcal{M}} + \left(\widetilde{X}\mathbf{M}\widetilde{X}\right)^{-1}\widetilde{X}\mathbf{M}\breve{\mathcal{E}}$.

$$\begin{aligned}\left|\mathbf{M}\widetilde{\mathcal{M}}\right| &\leqslant \left|\widetilde{\mathcal{M}} - \widetilde{P}\beta^{0,K}\right| + \left|\widetilde{P}\beta^{0,K} - \widetilde{P}\mathbf{b}^K\right|\\ &\leqslant \sqrt{NT}O_p\left(K^{-\mu} + \sqrt{K}\Pi_K N^{-1}\right) + \sqrt{NT}\xi_K O_p\left(K^{-\mu} + \sqrt{K}\Pi_K N^{-1}\right)\\ &= o(1).\end{aligned}$$

The second last line uses **Lemma 12**, Assumption 1, and Assumption 3. And the last line uses the rates specified in Assumption 9 and 5. Since $\frac{1}{\sqrt{NT}}\widetilde{X}'\mathbf{M} = O_p(1)$, therefore $\frac{1}{\sqrt{NT}}\widetilde{X}'\mathbf{M}\widetilde{\mathcal{M}} = o_p(1)$. Then adding **Lemma 14**,

$$\frac{1}{\sqrt{NT}}\left(\breve{\theta} - \theta^0\right) = \left(\sum_{g=1}^{G^0}\kappa_g\psi_g^{xz} + o_p(1)\right)^{-1}\frac{1}{\sqrt{NT}}\widetilde{X}'\mathbf{M}\mathcal{E} + o_p(1).$$

Note,

$$\begin{aligned}\cdot\left[\left|\widetilde{P}\left(\widetilde{P}'\widetilde{P}\right)^{-1}\widetilde{P}'\mathcal{E}\right|^2 \mid Z\right] &= tr\left(\widetilde{P}\left(\widetilde{P}'\widetilde{P}\right)^{-1}\widetilde{P}'\cdot\left[\mathcal{E}\mathcal{E}' \mid Z\right]\right)\\ &\lesssim M^{\frac{3}{2}}TK\end{aligned}$$

As $\cdot\left[\mathcal{E}\mathcal{E}' \mid Z\right]$ is block diagonal with $T \times T$ submatrices, an analogous **Lemma 14** argument provides the last line. Hence, $\frac{1}{\sqrt{NT}}\mathbf{M}\mathcal{E} = \frac{\mathcal{E}}{\sqrt{NT}} + O_p\left(\frac{\sqrt{K}}{\sqrt{N}}\right)$. Finally,

$$\frac{1}{\sqrt{NT}}\left(\breve{\theta} - \theta^0\right) = \left(\sum_{g=1}^{G^0}\kappa_g\psi_g^{xz} + o_p(1)\right)^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \cdot[x_{it} \mid z_{it}] - \psi^{x}(\alpha,t) + \cdot\left[\psi^{x}(\alpha,t) \mid z_{it}\right]\right)\epsilon_{it} + o_p(1).$$

From **Lemma 13**, $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\cdot\left[x_{it}|z_{it}\right]-\psi^{x}\left(\alpha,t\right)+\cdot\left[\psi^{x}\left(\alpha,t\right)|z_{it}\right]\right)\epsilon_{it}\Rightarrow N(0,\psi^{x\epsilon})$. Then using the continuous mapping theorem, $\sqrt{NT}\left(\check{\theta}-\theta^{0}\right)\Rightarrow N(0,\Sigma_{\theta})$. □

Lemma 18. *Under Assumption 1-6 and 8-9,* $\sqrt{NT}\left(\check{\zeta}-\tilde{\zeta}\right)=o_{p}\left(1\right)$.

Proof. By the least squares solution, $\left(\tilde{\theta},\tilde{\beta}^{K}\right)$ minimizes the criterion

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\overline{y_{g_{i}^{0}t}}-\left(x_{it}-\overline{x_{g_{i}^{0}t}}\right)'\theta-\left(p^{K}\left(z_{it}\right)-\overline{p_{g_{i}^{0}t}^{K}}\right)'\beta^{K}\right)^{2}$$

For the sake of this lemma's proof, redefine

$$\tilde{Q}_{c}:=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\overline{y_{g_{i}^{0}t}}-\left(x_{it}-\overline{x_{g_{i}^{0}t}}\right)'\theta-\left(p^{K}\left(z_{it}\right)-\overline{p_{g_{i}^{0}t}^{K}}\right)'\beta^{K}\right)^{2}.$$

Applying the mean-value theorem on the F.O.C from ζ^{0}:

1. $0=\sqrt{NT}\frac{\partial\tilde{Q}_{c}}{\partial\zeta}\left(\zeta^{0}\right)+\frac{\partial^{2}\tilde{Q}_{c}}{\partial\zeta\partial\zeta'}\left(\zeta^{0}\right)\sqrt{NT}\left(\tilde{\zeta}-\zeta^{0}\right)$. [3]
2. $0=\sqrt{NT}\frac{\partial\widetilde{Q}_{c}}{\partial\zeta}\left(\zeta^{0}\right)+\frac{\partial^{2}\widetilde{Q}_{c}}{\partial\zeta\partial\zeta'}\left(\zeta^{0}\right)\sqrt{NT}\left(\check{\zeta}-\zeta^{0}\right)$.

Thus if (a) $\frac{\partial\widetilde{Q}_{c}}{\partial\zeta}\left(\zeta^{0}\right)=\frac{\partial\tilde{Q}_{c}}{\partial\zeta}\left(\zeta^{0}\right)+o_{p}\left(\frac{1}{\sqrt{NT}}\right)$ and (b) $\frac{\partial^{2}\tilde{Q}_{c}}{\partial\zeta\partial\zeta'}\left(\zeta^{0}\right)=\frac{\partial^{2}\widetilde{Q}_{c}}{\partial\zeta\partial\zeta'}\left(\zeta^{0}\right)+o_{p}\left(1\right)$ are true, then the proof is complete.
On **part (a)**. Based on Assumption 1's approximation and Assumption 9's definition, WLOG $\left|\psi_{g}^{m}-\left(\psi_{g}^{p,K}\right)'\beta^{0,K}\right|=O\left(K^{-\mu}\right)$ and

$$\left|m-\left(p^{K}\right)'\beta^{0,K}\right|_{\mathscr{X},\mathscr{Z}}=O\left(K^{-\mu}\right).$$

Then $\sqrt{NT}\left[\frac{\partial\tilde{Q}_{c}}{\partial\zeta}\left(\zeta^{0}\right)-\frac{\partial\widetilde{Q}}{\partial\zeta}\left(\zeta^{0}\right)\right]=\underbrace{\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix}O\left(K^{-\mu}\right)\left(\psi_{g_{i}^{0}}^{x}\left(\alpha,t\right)-\overline{x_{g_{i}^{0}t}}\right)\\O\left(K^{-\mu}\right)\left(\psi_{g_{i}^{0}}^{p,K}\left(\alpha,t\right)-\overline{p_{g_{i}^{0}t}^{K}}\right)\end{pmatrix}}_{A_{1}}$
$+\underbrace{\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix}\left(x_{it}-\overline{x_{g_{i}^{0}t}}\right)\epsilon_{g_{i}^{0}t}\\\left(p^{K}\left(z_{it}\right)-\overline{p_{g_{i}^{0}t}^{K}}\right)\epsilon_{g_{i}^{0}t}\end{pmatrix}}_{A_{2}}-\underbrace{\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix}\left(\psi_{g_{i}^{0}}^{x}\left(\alpha,t\right)-\overline{x_{g_{i}^{0}t}}\right)\epsilon_{it}\\\left(\psi_{g_{i}^{0}}^{p,K}\left(\alpha,t\right)-\overline{p_{g_{i}^{0}t}^{K}}\right)\epsilon_{it}\end{pmatrix}}_{A_{3}}.$
Analyzing the terms,

- $\|A_{1}\|=o_{p}\left(1\right)$ as $\frac{\sqrt{NT}\xi_{K}}{K^{\mu}}\to0$.
- **on** A_{3}. A standard law of large number gives $\left|\frac{\sum_{i=1}^{N}\epsilon_{it}}{\sqrt{N}}\right|=O_{p}\left(1\right)$. Hence, $\sup_{1\leqslant t\leqslant T}\left|\frac{\sum_{i=1}^{N}\epsilon_{it}}{\sqrt{N}}\right|=O_{p}\left(T\right)$. Similarly, $\psi_{g_{i}^{0}}^{p,K}\left(\alpha,t\right)-\overline{p_{g_{i}^{0}t}^{K}}=O_{p}\left(\frac{\xi_{K}}{\sqrt{N}}\right)$ and $\psi_{g_{i}^{0}}^{x}-\overline{x_{g_{i}^{0}t}}=O_{p}\left(\frac{1}{\sqrt{N}}\right)$.
  Thus $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi_{g_{i}^{0}t}^{x}-\overline{x_{g_{i}^{0}t}}\right)\epsilon_{it}=\sum_{g=1}^{G^{0}}\sum_{t=1}^{T}\left[\frac{\psi_{g}^{p,K}\left(\alpha,t\right)-\overline{p_{g_{i}^{0}t}^{K}}}{\sqrt{T}}\left[\frac{\sum_{i=1}^{N}\epsilon_{it}\{g_{i}^{0}=g\}}{\sqrt{N}}\right]\right]=O_{p}\left(\frac{T^{\frac{3}{2}}}{\sqrt{N}}\right)$. And similarly,

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi_{g_{i}^{0}t}^{p,K}-\overline{p_{g_{i}^{0}t}^{K}}\right)\epsilon_{it}=O_{p}\left(\frac{\xi_{K}T^{\frac{3}{2}}}{\sqrt{N}}\right).$$

[3] Technically, the mean value theorem evaluates the Hessian at the midpoint between the two points. But evaluating the Hessian at ζ^{0} is not an issue because the Hessian is constant.

Then by rates in Assumption 5 and 9, $\|A_3\| = o_p(1)$.

- $\|A_2\| = o_p(1)$ by an analysis similar to the above with A_3.

On **part (b)**.

$$\frac{\partial^2 \bar{Q}_c}{(\partial\zeta)^2}(\zeta) = \frac{\partial^2 \overline{Q}_c}{(\partial\zeta)^2}(\zeta) + \underbrace{\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix}\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)\left(\left(x_{it} - \psi^x_{g_i^0}(\alpha,t)\right)', \left(p^K(z_{it}) - \psi^{p_{i^0}K}_{g_i^0}(\alpha,t)\right)'\right) \\ \left(\psi^x_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)\left(\left(x_{it} - \psi^x_{g_i^0}(\alpha,t)\right)', \left(p^K(z_{it}) - \psi^{p_{i^0}K}_{g_i^0}(\alpha,t)\right)'\right)\end{pmatrix}}_{A_4}$$

$$+ \underbrace{\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix}\left(x_{it} - \overline{x_{g_i^0 t}}\right)\left(\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)', \left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)'\right) \\ \left(p^K(z_{it}) - \overline{p^K_{g_i^0 t}}\right)\left(\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)', \left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)'\right)\end{pmatrix}}_{A_5}.$$

On A_5: From conditional independence, $\frac{1}{N}\sum_{i=1}^{N}\left(x_{it} - \overline{x_{g_i^0 t}}\right)\{g_i^0 = g\} = O_p\left(\frac{1}{\sqrt{N}}\right)$ and $\frac{1}{N}\sum_{i=1}^{N}\left(p^K(z_{it}) - \overline{p^K_{g_i^0 t}}\right)\{g_i^0 = g\} = O_p\left(\frac{\xi_K}{\sqrt{N}}\right)$. Then, from the stationary assumption, $\sup_{1\leqslant t\leqslant T}\left|\frac{1}{N}\sum_{i=1}^{N}\left(x_{it} - \overline{x_{g_i^0 t}}\right)\{g_i^0 = g\}\right| = O_p\left(\frac{T}{\sqrt{N}}\right)$ and $\sup_{1\leqslant t\leqslant T}\left|\frac{1}{N}\sum_{i=1}^{N}\left(p^K(z_{it}) - \overline{p^K_{g_i^0 t}}\right)\{g_i^0 = g\}\right| = O_p\left(\frac{\xi_K T}{\sqrt{N}}\right)$. Furthermore, $\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}} = O_p\left(\frac{1}{\sqrt{N}}\right)$, and $\psi^{p_{i^0}K}_{g_i^0}(\alpha.t) - \overline{p^K(z_{it})} = O_p\left(\frac{1}{\sqrt{N}}\right)$.

Notice, $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x_{g_i^0 t}}\right)\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)' = 2\sum_{g=1}^{G}\sum_{t=1}^{T}\left[\frac{\sum_{i=1}^{N}\left(x_{it} - \overline{x_{gt}}\right)\{g_i^0 = g\}}{N}\right]\frac{\left(\psi^x_g - \overline{x_{gt}}\right)'}{T} = O_p\left(\frac{T}{N}\right)$. By similar factoring, it can be shown that $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x_{g_i^0 t}}\right)\left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)' = O_p\left(\frac{\xi_K T}{N}\right)$,

$\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p(z_{it}) - \overline{p^K_{g_i^0 t}}\right)\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)' = O_p\left(\frac{\xi_K T}{N}\right)$, and $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(p(z_{it}) - \overline{p^K_{g_i^0 t}}\right)\left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)' = O_p\left(\frac{\xi_K^2 T}{N}\right)$.
Thus $A_5 = o(1)$, under the rates assumed in Assumption 5 and 9.
On A_4 :
Similar to A_5's reasoning, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)\left(x_{it} - \psi^x_{g_i^0 t}(\alpha,t)\right)' = O_p\left(\frac{T}{N}\right)$,
$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)\left(x_{it} - \psi^x_{g_i^0 t}(\alpha,t)\right)' = O_p\left(\frac{T\xi_K}{N}\right)$,
$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi^x_{g_i^0}(\alpha,t) - \overline{x_{g_i^0 t}}\right)\left(p(z_{it}) - \psi^{p_{i^0}K}_{g_i^0 t}(\alpha,t)\right)' = O_p\left(\frac{\xi_K T}{N}\right)$,
and $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\psi^{p_{i^0}K}_{g_i^0}(\alpha,t) - \overline{p^K_{g_i^0 t}}\right)\left(p^K(z_{it}) - \psi^{p_{i^0}K}_{g_i^0 t}(\alpha,t)\right)' = O_p\left(\frac{\xi_K^2 T}{N}\right)$. Hence, also $A_4 = O_p(1)$. □

Lemma 19. *Under Assumption 1-6 and 8-9, as* $N, T, K \to \infty$,

$$\left|m - \left(p^K\right)'\breve{\beta}^K\right|_{\infty,\mathcal{Z}} = O_p\left(\frac{\xi_K\sqrt{K}}{\sqrt{N}}\right) + O_p\left(\frac{\xi_K}{K^\mu}\right) + O_p\left(\frac{\xi_K\sqrt{K}\Pi_K}{N}\right).$$

Proof.

$$\begin{aligned}\breve{\beta}^K &= \left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\left(\breve{Y} - \breve{X}\breve{\theta}\right) \\ &= \left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\left(\breve{\mathcal{M}}\right) + \left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\mathcal{E} + \left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\breve{X}\left(\breve{\theta} - \theta^0\right)\end{aligned}$$

Note, $\cdot\left[\left|\breve{P}'\breve{X}\right|^2\right] = \cdot\left[tr\left(\breve{X}'\breve{P}\breve{P}'\breve{X}\right)\right] = NTtr\left(\cdot\left[\breve{X}'\breve{X}\right]\right) = O(NT)$. Then, $\cdot\left[\left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\breve{X}\left(\breve{\theta} - \theta^0\right)\right] = \cdot\left[\left(\frac{1}{NT}\breve{P}'\breve{P}\right)^{-1}\frac{1}{NT}\breve{P}'\breve{X}\left(\breve{\theta} - \theta^0\right)\right] = O\left(\frac{1}{\sqrt{NT}}\right)$.

From **Lemma 12**, $\left(\breve{P}'\breve{P}\right)^{-1}\breve{P}'\left(\breve{\mathcal{M}}\right) = \beta^{0,K} + O_p\left(K^{-\mu} + \sqrt{K}\Pi_K N^{-1}\right)$.
In summation form,

$$\begin{aligned}\cdot\left[\left|\frac{1}{N}\breve{P}'\mathcal{E}\right|^2\right] &= \sum_{l=1}^{K}\cdot\left[\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{p}_l^K(z_{it})\epsilon_{it}\right)^2\right] \\ &= \sum_{l=1}^{K}\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}Cov\left(\mathbf{p}_l^K(z_{it})\epsilon_{it}, \mathbf{p}_l^K(z_{is})\epsilon_{is}\right) \\ &= \sum_{l=1}^{K}\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{t=1}^{T}Cov\left(\mathbf{p}_l^K(z_{i1})\epsilon_{i1}, \mathbf{p}_l^K(z_{is})\epsilon_{is}\right) \\ &\lesssim O\left(\frac{K\xi_K^2}{NT}\right)\end{aligned}$$

The second line applies the conditional independence and the third line applies the stationary assumption. By using Davydov inequality and the summed mixing coefficients as convergent, repeating the **Lemma 11**'s argument delivers the last line. Hence, by Jensen inequality, $\left|\frac{1}{N}\bar{P}'\mathcal{E}\right| = O_p\left(\frac{\sqrt{K}\xi_K}{\sqrt{NT}}\right)$ and, with **Lemma 11**, $\left(\bar{P}'\bar{P}\right)^{-1}\bar{P}'\mathcal{E} = O_p\left(\frac{\sqrt{K}\xi_K}{\sqrt{NT}}\right)$.
Then,

$$\begin{aligned}\left|m - \left(p^K\right)'\check{\beta}^K\right|_{\infty,\mathcal{Z}} &\leqslant \left|m - \left(p^K\right)'\beta^{0,K}\right|_{\infty,\mathcal{Z}} + \left|\left(p^K\right)'\left(\tilde{\beta} - \check{\beta}\right)\right|_{\infty,\mathcal{Z}} + \left|\left(p^K\right)'\left(\beta^{0,K} - \check{\beta}\right)\right|_{\infty,\mathcal{Z}} \\ &\leqslant O_p\left(K^{-\mu}\right) + O_p\left(\xi_K K^{-\mu} + \xi_K\sqrt{K}\Pi_K N^{-1}\right) + O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right) + O_p\left(\frac{\xi_K}{\sqrt{NT}}\right) \\ &= O_p\left(\xi_K K^{-\mu} + \xi_K\sqrt{K}\Pi_K N^{-1}\right) + O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right)\end{aligned}$$

□

Lemma 20. *Under Assumption 1-6 and 8-9, as $N, T, K \to \infty$,*

1. $\sqrt{NT}\left(\tilde{\theta} - \theta^0\right) \Rightarrow N(0, \Sigma_\theta)$, *and*
2. *For any $g \in \{1, ..., G^0\}$,* $\sup_{t\in\{1,...,T\}}\left|\tilde{\alpha}_{g^0 t} - \alpha^0_{g^0 t}\right| = o_p(1)$.

Proof. Point 1 follows immediately from combining **Lemma 15**, $\sqrt{NT}\left(\check{\theta} - \theta^0\right) \Rightarrow N(0, \Sigma_\theta)$, with **Lemma 16**, $\sqrt{NT}\left(\check{\theta} - \tilde{\theta}\right) = o_p(1)$.
To prove point 2:

$$\begin{aligned}\cdot\left[\sup_{t\in\{1,...,T\}}\left|\tilde{\alpha}_{g^0 t} - \alpha^0_{g^0 t}\right|\right] &\leqslant \sum_{t=1}^{T}\cdot\left[\left|\tilde{\alpha}_{g^0 t} - \alpha^0_{g^0 t}\right|\right] \\ &\leqslant T\cdot\left[\left|\tilde{\alpha}_{g^0 1} - \alpha^0_{g^0 1}\right|\right] \\ &\leqslant \frac{T}{N^{\delta'}}\left(\cdot\left[\left|\overline{x_{g^0 t}}\right|\left|N^{\delta'}\left(\tilde{\theta} - \theta^0\right)\right|\right] + N^{\delta'}\cdot\left[\left|m - \left(p^K\right)'\tilde{\beta}^K\right|^2_{\infty,\mathcal{Z}}\right]^{\frac{1}{2}} + M\cdot\left[\left|N^{\delta'}\overline{u_{g^0 t}}\right|\right]\right)\end{aligned}$$

The first inequality follows from the maximum is less than the total sum. The second inequality follows from the stationary assumption. And the third inequality follows from $\tilde{\alpha}_{g^0 t}$ being a least squares solution, the triangle inequality, and Assumption 3.

From the standard Lindeberg and Feller central limit theorem, $\sqrt{N}\overline{u_{g^0 t}} = O_p(1)$. Therefore, $\cdot\left[\left|N^{\delta'}\overline{u_{g^0 t}}\right|\right] = o(1)$, as $\delta' < \frac{1}{2}$.

From Assumption 3's moment bounds, $\left|\overline{x_{g^0 t}}\right| = O_p(1)$ and, from point 1, $\sqrt{N^{\delta'}}\left(\tilde{\theta} - \theta^0\right) = o_p(1)$, as $\delta' < \frac{1}{2}$. Therefore, by continuous mapping, $\cdot\left[\left|\overline{x_{g^0 t}}\right|\left|N^{\delta'}\left(\tilde{\theta} - \theta^0\right)\right|\right] = o(1)$.

Similarly, from $\delta' < \frac{1}{2}$, the previous lemma and Assumption 9's rate gives $N^{\delta'}\cdot\left[\left|m - \left(p^K\right)'\tilde{\beta}^K\right|^2_{\infty,\mathcal{Z}}\right]^{\frac{1}{2}} = O(1)$.

In conclusion, by Markov inequality, $\sup_{t\in\{1,...,T\}}\left|\tilde{\alpha}_{g^0 t} - \alpha^0_{g^0 t}\right| = o_p(1)$. □

Proof of Theorem 4. From **Lemma 9** and **Lemma 18**, $\sqrt{NT}\left(\hat{\theta} - \theta^0\right) = \sqrt{NT}\left(\hat{\theta} - \tilde{\theta}\right) + \sqrt{NT}\left(\tilde{\theta} - \theta^0\right) \Rightarrow N(0, \Sigma_\theta)$. Also from **Lemma 17** and **Lemma 16**, $\|\hat{m} - m\|_{\infty,\mathcal{Z}} \leqslant \left|m - \left(p^K\right)\tilde{\beta}\right|_{\infty,\mathcal{Z}} + \xi_K\left|\hat{\beta}^K - \tilde{\beta}^K\right|$
$= O_p\left(\xi_K K^{-\mu} + \xi_K\sqrt{K}\Pi_K N^{-1}\right) + O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right) + O_p\left(\frac{\xi_K}{\sqrt{NT}}\right)$
$= O_p\left(\xi_K K^{-\mu} + \xi_K\sqrt{K}\Pi_K N^{-1}\right) + O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right)$.
Now to prove the uniform convergence part.

$$\begin{aligned}\sup_{i\in\{1,...,N\};t\in\{1,...,T\}}\left|\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t}\right| &\leqslant \sup_{i\in\{1,...,N\};t\in\{1,...,T\}}\left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| + \sup_{i\in\{1,...,N\};t\in\{1,...,T\}}\left|\tilde{\alpha}_{g_i^0 t} - \alpha^0_{g_i^0 t}\right| \\ &\leqslant \sup_{i\in\{1,...,N\};t\in\{1,...,T\}}\left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| + \sum_{g=1}^{G^0}\sup_{t\in\{1,...,T\}}\left|\tilde{\alpha}_{gt} - \alpha^0_{gt}\right|\end{aligned}$$

The first line comes from triangle inequality. There are only G^0 different α estimates at every period t. This fact leads to the second inequality.

From **Lemma 18**, $\sum_{g=1}^{G^0} \sup_{t\in\{1,\dots,T\}} \left|\tilde{\alpha}_{gt} - \alpha^0_{gt}\right| = o_p(1)$.

$$\cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right|\right] \leqslant \cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| \left\{\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| = 0\right\}\right] + \cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| \left\{\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| > 0\right\}\right].$$

From Θ's compactness, $\sup_{\alpha\in\Theta} |\alpha|$ is finite.

$$\begin{aligned}\cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| \left\{\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| > 0\right\}\right] &\leqslant 2\sup_{\alpha\in\Theta}|\alpha| \cdot \left(\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| > 0\right) \\ &= o(1)\end{aligned}$$

The last equality follows from Theorem 3: $\cdot\left(\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| > 0\right) = o(1)$.

$$\begin{aligned}\cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| \left\{\sup_{i\in\{1,\dots,N\}} \left|\hat{g}_i - \phi\left(g_i^0\right)\right| = 0\right\}\right] &\leqslant \cdot\left[\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\phi\left(g_i^0\right)t} - \tilde{\alpha}_{g_i^0 t}\right|\right] \\ &\leqslant \sum_{g=1}^{G^0} \cdot\left[\sup_{t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\phi(g)t} - \tilde{\alpha}_{gt}\right|\right] \\ &= o(1)\end{aligned}$$

Again, the second inequality comes from α's estimate differs only by the group. The last equality follows from **Lemma 9**. In particular, $\sup_{t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\phi(g)t} - \tilde{\alpha}_{gt}\right| \leqslant \sqrt{\sum_{t=1}^{T} \left(\hat{\alpha}_{\phi(g)t} - \tilde{\alpha}_{gt}\right)^2} = o_p\left(T^{\frac{1-\delta}{2}}\right)$ and $\delta > 1$ from Assumption 8.

Therefore, Markov inequality gives $\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \tilde{\alpha}_{g_i^0 t}\right| = o_p(1)$.

Finally, $\sup_{i\in\{1,\dots,N\};t\in\{1,\dots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t}\right| = o_p(1)$.

□

Proof of Corollary 2. By **Theorem 3**, it is sufficient to prove $\hat{\Sigma}_\theta$ as consistent under the Oracle condition, i.e. by assuming $\hat{g}_i = g_i^0$. Stacking $x_{it} - \frac{\sum_{j:g_j^0=g_i^0} x_{jt}}{N_{g_i^0}}$, $\hat{\epsilon}_{it}$, $p^K(z_{it}) - \frac{\sum_{j:g_j^0=g_i^0} p^K\left(z_{jt}\right)}{N_{g_i^0}}$ to form matrices $\tilde{X}$, $\tilde{P}$, and $\tilde{\mathcal{E}}$, respectively. Then define $\tilde{M} = I_{NT} - \tilde{P}\left(\tilde{P}'\tilde{P}\right)^{-1}\tilde{P}'$.

Under the Oracle condition, $\sum_{g=1}^{G^0} \frac{\tilde{N}_g}{N} \hat{\psi}^{xz} = \frac{1}{NT}\tilde{X}'\tilde{M}\tilde{X}$ and $\hat{\psi}^{x\epsilon} = \frac{1}{NT}\sum_{i=1}^{N}\left[\tilde{X}'\tilde{M}\tilde{\mathcal{E}}\right]_i\left[\tilde{\mathcal{E}}'\tilde{M}\tilde{X}\right]_i$, where the subscript i denotes the submatrix containing only the ith unit's observations.

If **(1)** $\frac{1}{NT}\tilde{X}'\tilde{M}\tilde{X} = \sum_{g=1}^{G^0}\kappa_g\psi_g^{xz} + o_p(1)$ and **(2)** $\frac{1}{NT}\sum_{i=1}^{N}\left[\tilde{X}'\tilde{M}\tilde{\mathcal{E}}\right]_i\left[\tilde{\mathcal{E}}'\tilde{M}\tilde{X}\right]_i = \psi^{x\epsilon} + o_p(1)$ hold then the continuous mapping theorem implies consistency of $\hat{\Sigma}_\theta$.

On (1):

By Markov inequality, conditional independence, and Assumption 9's bound, $\frac{\sum_{j:g_j^0=g_i^0} x_{jt}}{N_{g_i^0}} = \psi^x_{g_i^0}(\alpha,t) + O_p\left(\frac{1}{\sqrt{N}}\right)$ and $\frac{\sum_{j:g_j^0=g_i^0} p^K\left(z_{jt}\right)}{N_{g_i^0}} = \psi^{p^K}_{g_i^0}(\alpha,t) + O_p\left(\frac{\xi_K}{\sqrt{N}}\right)$. Hence, $\frac{1}{\sqrt{NT}}\left|\tilde{X} - \overline{X}\right| = O_p\left(\frac{1}{\sqrt{N}}\right)$ and $\frac{1}{\sqrt{NT}}\left|\tilde{P} - \overline{P}\right| = O_p\left(\frac{\xi_K}{\sqrt{N}}\right)$. Furthermore, $\frac{1}{\sqrt{NT}}\left|\left(\tilde{P}'\tilde{P}\right) - \left(\overline{P}'\overline{P}\right)\right| \leqslant \frac{1}{\sqrt{NT}}\left|\left(\tilde{P} - \overline{P}\right)'\tilde{P}\right|$
$+ \frac{1}{\sqrt{NT}}\left|\overline{P}'\left(\tilde{P} - \overline{P}\right)\right| = O_p\left(\frac{\xi_K^2}{\sqrt{N}}\right)$. Similarly, $\frac{1}{NT}\left|\left(\tilde{P}\tilde{P}'\right) - \left(\overline{P}\overline{P}'\right)\right| = O_p\left(\frac{\xi_K^2}{\sqrt{N}}\right)$.

Thus $\frac{1}{\sqrt{NT}}\left(\mathbf{M} - \tilde{M}\right) = O_p\left(\frac{\xi_K^2}{\sqrt{N}}\right)$.

Furthermore, the decomposition: $\frac{1}{NT}\tilde{X}'\tilde{M}\tilde{X} = \frac{1}{NT}\overline{X}'\mathbf{M}\overline{X} + \frac{1}{\sqrt{NT}}\left(\tilde{X} - \overline{X}\right)'\frac{1}{\sqrt{NT}}\left(\mathbf{M}\overline{X}\right)$
$+ \tilde{X}'\frac{1}{NT}\left(\tilde{M} - \mathbf{M}\right)\overline{X} + \frac{1}{\sqrt{NT}}\left(\tilde{X}\tilde{M}\right)\frac{1}{\sqrt{NT}}\left(\tilde{X} - \overline{X}\right)$ implies $\frac{1}{NT}\tilde{X}'\tilde{M}\tilde{X} = \frac{1}{NT}\overline{X}'\mathbf{M}\overline{X} + o_p(1)$, under the provided rates.

Then applying **Lemma 14** gives $\frac{1}{NT}\tilde{X}'\tilde{M}\tilde{X} = \sum_{g=1}^{G^0}\kappa_g\psi_g^{xz} + o_p(1)$.

On (2):

Let $\tilde{\beta}_x = \left(\tilde{P}'\tilde{P}\right)^{-1}\tilde{P}'\tilde{X}$, then above observations show $\tilde{\beta}_x = \left(\overline{P}'\overline{P}\right)^{-1}\overline{P}'\overline{X} + o_p(1)$. Furthermore, adding **Lemma 14**'s argument implies

$\tilde{\beta}_x = (\overline{P}'\overline{P})^{-1}\overline{P}'_{\cdot Z}[\overline{X}] + o_p(1) = \mathbf{b}_x^K + o_p(1)$. Thus,

$$\begin{aligned}
\ddot{x}_{it,j} &= x_{it} - \frac{\Sigma_{j:g_j^0=g}\, x_{jt}}{N_{g_i^0}} - \left(p^K(z_{it}) - \frac{\Sigma_{j:g_j^0=g}\, p^K(z_{jt})}{N_{g_i^0}}\right)'\tilde{\beta}_{x,j} \\
&= \left(\mathbf{v}_j^{g_i^0}(z_{it}) - \mathbf{p}^K(z_{it})'\mathbf{b}_{x,j}^K\right) + \left(x_{it} - \cdot[x_{it}|z_{it}] - \psi_{g_i^0}^x(\alpha,t) + \cdot\left[\psi_{g_i^0}^x(\alpha,t)|z_{it}\right]\right) + \left(\psi_{g_i^0}^x(\alpha,t) - \frac{\Sigma_{j:g_j^0=g}\, x_{jt}}{N_{g_i^0}}\right) \\
&+ \left(\psi_{g_i^0}^{p^K}(\alpha,t) - \frac{\Sigma_{j:g_j^0=g}\, p^K(z_{jt})}{N_{g_i^0}}\right)'\mathbf{b}_{x,j}^K + o_p(1) \\
&= x_{it} - \cdot[x_{it}|z_{it}] - \psi_{g_i^0}^x(\alpha,t) + \cdot\left[\psi_{g_i^0}^x(\alpha,t)|z_{it}\right]
\end{aligned}$$

And Theorem 4 provides $\sup_{i,t}\cdot[|\hat{\epsilon}_{it} - \epsilon_{it}|] = o_p(1)$. Hence, $\hat{\psi}^{x\epsilon} = \frac{\Sigma_{i=1}^N \Sigma_{t=1}^T \Sigma_{s=1}^T w_{it}w_{is}}{NT} + o_p(1)$, where w_{it} is defined in **Lemma 13**. Finally, like in previous arguments, the conditional independence, stationary property and Davydov inequality lead to $\hat{\psi}^{x\epsilon} = \psi^{x\epsilon} + o_p(1)$. □

Appendix - Chapter 3

Section 3.1 The corresponding population criterion of $\hat{\mathfrak{m}}_{NT}$ is

$$\mathfrak{M}(\tau, \hat{\mathfrak{h}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \hat{\mathfrak{h}}\left(w_{it}, z_{it}, \tau\right)\right)\right]' W \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \hat{\mathfrak{h}}\left(w_{it}, z_{it}, \tau\right)\right)\right].$$

Chen, Linton, and Keilegom (2003)'s Theorem 1 for consistency applies here by replacing their sequence index n with NT. Their argument uses assumptions on the criterion function's behavior and not the data generating process. Proof of Theorem S 1 involves checking the assumptions used in Chen, Linton, and Keilegom (2003)'s Theorem 1. The following notations are helpful for proofs.

Notation:

1. $\mathfrak{z} := \{m^0(z) \,|\, z \in \mathcal{Z}\}$.

2. $\mathfrak{R}(\mathfrak{W} \times \mathcal{T}) := \{R^0(w, \tau) \,|\, w \in \mathfrak{W}; \tau \in \mathcal{T}\}$.

Lemma S 1. *Under Assumption S 1, for any $\delta > 0$ there exists a $\epsilon(\delta) > 0$ such that* $\sup_{\|\tau - \tau_0\| > \delta} \mathfrak{M}\left(\tau, \mathfrak{h}^0\right) > \epsilon(\delta)$.

Proof. Let $\lambda_{\min}$ be the smallest eigenvalue of W. As W is positive definite, $\lambda_{\min}$ is positive. Using Spectral Decomposition on W,

$$\mathfrak{M}\left(\tau, \mathfrak{h}^0\right) \geqslant \lambda_{\min} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^0\left(w_{it}, z_{it}, \tau\right)\right)\right]' \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^0\left(w_{it}, z_{it}, \tau\right)\right)\right]\right].$$

Let $\delta > 0$. Take the $\epsilon(\delta)'$ from Assumption S 1, then it follows that

$$\mathfrak{M}\left(\tau, \mathfrak{h}^0\right) > \lambda_{\min} \epsilon(\delta)',$$

whenever $\left|\tau - \tau^0\right| > \delta$. Hence, the conclusion follows by taking $\epsilon(\delta) = \lambda_{\min} \epsilon'(\delta)$. □

Lemma S 2. *Under Assumption 2, S 1, S 2, S 3, and S 4, there exists a compact set $\mathfrak{J} \subset \mathbb{R}^{2+d_2+d_3+d_4+d_5}$ and constants $\left(\delta_J, C^J\right)$ such that, for any $\left(w, \theta, m, \alpha_{gt}, \tau, R\right) \in \mathfrak{W} \times \Theta \times \mathfrak{z} \times \mathcal{A} \times \mathcal{T} \times \mathfrak{R}$, $\prod_{i \in \{w, \theta, m, \alpha_{gt}, \tau, R\}} B(i, \delta_J)$ is in the interior of $\mathfrak{J}$ and*

$$\begin{aligned}
&\left|\mathfrak{m}\left(w, \theta, m, \alpha_{gt}, \tau, R\right) - \mathfrak{m}\left(w, \theta', m', \alpha'_{gt}, \tau', R'\right)\right| \\
&\leqslant C^J \left[\left|\theta - \theta'\right| + \left|m - m'\right| + \left|\alpha'_{gt} - \alpha_{gt}\right| + \sum_{j=1}^{d_3} \left|R_j - R'_j\right| + \left|\tau - \tau'\right|\right],
\end{aligned}$$

for $\left(w,\theta,m,\alpha_{gt},\tau,R\right),\left(w,\theta',m',\alpha'_{gt},\tau',R'\right)\in\mathfrak{J}$.

Proof. By Assumption 2, S 2, and S 4, $\mathfrak{w},\Theta,\mathfrak{z},\mathcal{A},\mathcal{T}$, and $\mathfrak{R}$ are compact sets. Thus it is possible to pick a larger compact sets $(\mathfrak{w}',\Theta',\mathfrak{z}',\mathcal{A}',\mathcal{T}',$ and $\mathfrak{R}')$ to contain them in their interiors, i.e $\mathfrak{w}\subset\left(\mathfrak{w}'\right)^{o},\Theta\subset\left(\Theta'\right)^{o},\mathfrak{z}\subset\left(\mathfrak{z}'\right)^{o},\mathcal{A}\subset\left(\mathcal{A}'\right)^{o},\mathcal{T}\subset\left(\mathcal{T}'\right)^{o}$, and $\mathfrak{R}\subset\left(\mathfrak{R}'\right)^{o}$. Furthermore, choose an uniform δ_J as defined in the Lemma S 2 to fit the neighborhoods inside these larger compact sets' interior. Finally, refine $\mathfrak{R}'$ to be large enough to contain $\mathfrak{R}\left(\mathfrak{w}'\times\mathcal{T}'\right)$ in its interior with the same δ_J.

Let $\mathfrak{J}:=\mathfrak{w}'\times\Theta'\times\mathfrak{z}'\times\mathcal{A}'\times\mathcal{T}'\times\mathfrak{R}'$. Then $\mathfrak{J}$ is a compact set by Tychonoff's theorem. Furthermore, $\prod_{i\in\{w,\theta,m,\alpha_{gt},\tau,R\}}B\left(i,\delta_J\right)$ is in the interior of $\mathfrak{J}$, whenever $\{w,\theta,m,\alpha_{gt},\tau,R\}\in\mathfrak{w}\times\Theta\times\mathfrak{z}\times\mathcal{A}\times\mathcal{T}\times\mathfrak{R}$.
By Assumption S 4.2, $\mathfrak{m}_l$ is continuously differentiable on $\mathfrak{J}$, for $l=1,...,L$. Then, from applying the mean-value theorem, $\mathfrak{m}_l$ is Lipschitz over $\mathfrak{J}$, for $l=1,...,L$. Pick an uniform Lipschitz constant C^J over all $\mathfrak{m}_l$, for $l=1,...,L$. Thus the conclusion follows because $\mathfrak{m}$ is entry-wise Lipschitz by C^J. □

Lemma S 3. *Let $\epsilon>0$ and $\epsilon\geqslant 0$. Under Assumption 2, S 1, S 2, S 3, and S4,*

$$T^{\epsilon}\left|\mathfrak{m}\left(w,\theta,m\left(z\right),\alpha_{gt},\tau,R\left(w,\tau\right)\right)-\mathfrak{m}\left(w,\theta^0,m^0\left(z\right),\alpha_{gt}^0,\tau',R^0\left(w,\tau'\right)\right)\right|\leqslant\epsilon+C^J\left[\left(1+c^{R^0}d_3\right)T^{\epsilon}\left|\tau-\tau'\right|\right],$$

when $d\left(\mathfrak{h},\mathfrak{h}^0\right)<\min\{\frac{\epsilon}{C^J2d_3T^{\epsilon}},\delta_J\}$, *for* $(w,z,\tau),\left(w,z,\tau'\right)\in\mathfrak{w}'\times\mathcal{Z}\times\mathcal{T}'$.

Proof. From Assumption S 4.2, R^0 is continuously differentiable on $\mathfrak{w}'\times\mathcal{T}'$. Thus by mean value theorem, R^0 is Lipschitz on $\mathfrak{w}'\times\mathcal{T}'$ in each of its component. Hence, there exists an uniform Lipschitz constant c^{R^0} for R_j^0 on $\mathfrak{w}'\times\mathcal{T}'$, for $j=1,..,L$. Therefore, $\sum_{j=1}^{d_3}|R_j^0\left(w,\tau\right)-R_j^0\left(w,\tau\right)|\leqslant d_3c^{R^0}\left|\tau-\tau'\right|$, for $(w,\tau),\left(w,\tau'\right)\in\mathfrak{w}'\times\mathcal{T}'$.

Applying triangle inequality,

$$\begin{aligned}\sum_{j=1}^{d_3}\left|R_j\left(w,\tau\right)-R_j^0\left(w,\tau'\right)\right|&\leqslant\sum_{j=1}^{d_3}\left|R_j^0\left(w,\tau\right)-R_j\left(w,\tau\right)\right|+\sum_{j=1}^{d_3}\left|R_j^0\left(w,\tau\right)-R_j^0\left(w,\tau'\right)\right|\\&\leqslant d_3\sup_{w\in\mathfrak{w}}\sup_{\tau\in\mathcal{T}}\left|R^0\left(w,\tau\right)-R\left(w,\tau\right)\right|+\sum_{j=1}^{d_3}\left|R_j^0\left(w,\tau\right)-R_j^0\left(w,\tau'\right)\right|\\&\leqslant\frac{\epsilon}{2C_J}+d_3c^{R^0}\left|\tau-\tau'\right|\end{aligned}$$

The last line uses the uniform norm's bound and the above Lipschitz condition.

When $d\left(\mathfrak{h},\mathfrak{h}^0\right)<\delta_J$ and $(w,z,\tau),\left(w,z,\tau'\right)\in\mathfrak{w}'\times\mathcal{Z}\times\mathcal{T}'$, Lemma S 2 gives

$$\begin{aligned}&T^{\epsilon}\left|\mathfrak{m}\left(w,\theta,m\left(z\right),\alpha_{gt},\tau,R\left(w,\tau\right)\right)-\mathfrak{m}\left(w,\theta^0,m^0\left(z\right),\alpha_{gt}^0,\tau',R^0\left(w,\tau'\right)\right)\right|\\&\leqslant T^{\epsilon}C^J\left[\left|\theta-\theta^0\right|+\left|m\left(z\right)-m^0\left(z\right)\right|+\left|\alpha_{gt}^0-\alpha_{gt}\right|+\sum_{j=1}^{d_3}\left|R_j\left(w,\tau\right)-R_j^0\left(w,\tau'\right)\right|+\left|\tau-\tau'\right|\right]\\&\leqslant\epsilon+C^J\left(d_3c^{R^0}+1\right)T^{\epsilon}\left|\tau-\tau'\right|\end{aligned}$$

The second inequality uses the fact, $d\left(\mathfrak{h},\mathfrak{h}^0\right)<\frac{\epsilon}{2T^{\epsilon}C^Jd_3}$. □

Lemma S 4. *Suppose $\nu_{NT}=O\left(T^{-\epsilon}\right)$ is a positive sequence as $T\to\infty$ and for some $\epsilon\in[0,\frac{1}{2})$. Under Assumption 2, S 1, S 2, S 3, and S 4,*

$$\sup_{i\in\{1,...,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)\right]\right]\right|=o_p\left(T^{-\epsilon}\right),$$

as $N,T\to\infty$ and $\frac{\log(N)}{T^{1-2\epsilon}}\to 0$.

Proof. Let $\epsilon > 0$ and, then, set $\delta(\epsilon) := \min\left\{\frac{\epsilon}{6C^J\left(1+c^{R^0}d_3\right)T^\epsilon}, \delta_J\right\}$. As $\mathcal{T}$ is compact, the open cover $\{B(\tau, \delta(\epsilon))\}_{\tau\in\mathcal{T}}$ admits a finite subcover $\left\{B\left(\tau_j, \delta(\epsilon)\right)\right\}_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)}$.

Let $\tau \in \mathcal{T}$, then there exists a $\tau_{j*} \in \left\{1, ..., k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)\right\}$ such that $\tau \in B\left(\tau_{j*}, \delta(\epsilon)\right)$.

Conditional on $d\left(\mathfrak{h}, \mathfrak{h}^0\right) \leqslant \min\{\delta_J, \frac{\epsilon}{6C^J d_3 T^\epsilon}\}$,

$$
\begin{aligned}
&T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)\right]\right]\right| \\
&\leqslant T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]\right]\right| \\
&+T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)-\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]\right| \\
&+T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]\right]\right| \\
&\leqslant 2\left(\frac{\epsilon}{3}+C^J d_3 c^{R^0}\left|\tau-\tau_{j*}\right|\right)+T^{\kappa'}\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]\right]\right| \\
&\leqslant \epsilon+T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_{j*},\mathfrak{h}^0\left(w_{it},z_{it},\tau_{j*}\right)\right)\right]\right]\right|
\end{aligned}
$$

The second last line uses Lemma S 3. As τ is arbitrary,

$$
\begin{aligned}
&\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\min\{\delta_J,\frac{\epsilon}{6C^J d_3}\}} T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)\right]\right]\right| \\
&\leqslant \epsilon+\sup_{i\in\{1,\ldots,N\}}\max_{j=1,\ldots,k_\tau}\left\{T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right|\right\} \\
&\leqslant \epsilon+\sum_{i=1}^{N}\sum_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)} T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right| \\
&\leqslant \epsilon+\sum_{i=1}^{N}\sum_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)}\sum_{l=1}^{L} T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right|
\end{aligned}
$$

Now to show the sum is $o_p(1)$, as $N, T \to \infty$. To start, let $\epsilon' > 0$.

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{N}\sum_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)}\sum_{l=1}^{L} T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right| > \epsilon'\right) \\
&\leqslant \sum_{i=1}^{N}\sum_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)}\sum_{l=1}^{L}\mathbb{P}\left(T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right| > \epsilon'\right)
\end{aligned}
$$

Each $\mathfrak{m}_l$ is bounded on the compact set $\mathfrak{Z}$. Denote the uniform bound as $M^{\mathfrak{m}}$, i.e. $\sup_{\tau\in\mathcal{T}}\sup_{w\in\mathfrak{w}}\sup_{z\in\mathcal{Z}}\left|\mathfrak{m}_l\left(w,\tau,\mathfrak{h}^0(w,z,\tau)\right)\right| \leqslant M^{\mathfrak{m}}$. Then with Assumption S 3.1, Su, Shi, and Philips (2016)'s Lemma S1.1 provides

$$
\begin{aligned}
&\mathbb{P}\left(T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right| > \epsilon'\right) \\
&= O\left(\exp\left(-\frac{v_1\left(\epsilon'\right)^2}{v_0 T^{2\epsilon-1}+4M^{2\mathfrak{m}}T^{-2(1-\epsilon)}+\epsilon' 2M^{\mathfrak{m}}\left(\log(T)\right)^2 T^{-(1-\epsilon)}}\right)\right),
\end{aligned}
$$

for some constant $v_1, v_2 > 0$. Furthermore, $\mathfrak{m}_l$ being bounded allows the proof to choose v_1 and v_2 independent of i, as done so in the proof of Su, Shi, and Philips (2016)'s Lemma S1.2.

As in the proof of Ai and Chen (2003)'s Lemma 1, $k_\tau\left(\frac{\epsilon}{T^\epsilon}\right) \lesssim \frac{T^{d_3\epsilon}}{\epsilon^{d_3}}$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{k_\tau\left(\frac{\epsilon}{T^\epsilon}\right)}\sum_{l=1}^{L}\mathbb{P}\left(T^\epsilon\left|\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)-\mathbb{E}\left[\mathfrak{m}_l\left(w_{it},\tau_j,\mathfrak{h}^0\left(w_{it},z_{it},\tau_j\right)\right)\right]\right]\right|>\epsilon'\right)\\
&\leqslant O\left(\exp\left(-\frac{v_1\left(\epsilon'\right)^2}{v_0T^{2\epsilon-1}+4M^{2\mathfrak{m}}T^{-2(1-\epsilon)}+\epsilon'2M^{\mathfrak{m}}\left(\log\left(T\right)\right)^2T^{-(1-\epsilon)}}+\epsilon d_3\log\left(T\right)+\log\left(N\right)\right)\right)
\end{aligned}$$

As $\epsilon<\frac{1}{2}$, by the l'Hopital rule: $\lim_{T\to\infty}\epsilon d_3\log\left(T\right)\left(v_0T^{2\epsilon-1}+4M^{2\mathfrak{m}}T^{-2(1-\epsilon)}+\epsilon'2M^{\mathfrak{m}}\left(\log\left(T\right)\right)^2T^{-(1-\epsilon)}\right)=0$. With $\frac{\log\left(N\right)}{T^{1-2\epsilon}}\to 0$ and l'Hopital rule, $\log\left(N\right)\left(v_0T^{2\epsilon-1}+4M^{2\mathfrak{m}}T^{-2(1-\epsilon)}+\epsilon'2M^{\mathfrak{m}}\left(\log\left(T\right)\right)^2T^{-(1-\epsilon)}\right)$
$=v_0\frac{\log\left(N\right)}{T^{1-2\epsilon}}+4M^{2\mathfrak{m}}\frac{\log\left(N\right)}{T^{2-2\epsilon}}+\epsilon'2M^{\mathfrak{m}}\frac{\log\left(N\right)}{T^{1-2\epsilon}}\left(\frac{\log\left(T\right)}{T^{\frac{1}{2}}}\right)^2=o\left(1\right)$. The proof concludes by the fact that both ϵ and ϵ' are arbitrary. □

Lemma S 5. *Under Assumption S 1, S 2, S 3, S 4, and S 5,*

$$\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left|\hat{\mathfrak{M}}_{NT}\left(\tau,\mathfrak{h}\right)-\mathfrak{M}\left(\tau,\mathfrak{h}\right)\right|=o_p\left(1\right),$$

as $N,T\to\infty$.

Proof. Define $\tilde{m}_{iT}\left(\tau,\mathfrak{h}\right):=\frac{1}{T}\sum_{t=1}^{T}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\left(w_{it},z_{it},\tau\right)\right)\right]\right]$.
Applying the triangle inequality and Cauchy-Schwartz inequality,

$$\begin{aligned}
&\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left|\hat{\mathfrak{M}}_{NT}\left(\tau,\mathfrak{h}\right)-\mathfrak{M}\left(\tau,\mathfrak{h}\right)\right|\\
&\leqslant\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|W\right\|\left|\tilde{m}_{iT}\left(\tau,\mathfrak{h}\right)\right|^2+4\frac{1}{N}\sum_{i=1}^{N}\left|\tilde{m}_{iT}\left(\tau,\mathfrak{h}\right)\right|\left\|W\right\|LM^{\mathfrak{m}}\right]\\
&\leqslant\left\|W\right\|\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left|\tilde{m}_{iT}\left(\tau,\mathfrak{h}\right)\right|^2+4\left\|W\right\|LM^{\mathfrak{m}}\left\|W\right\|\sup_{i\in\{1,\ldots,N\}}\sup_{\tau\in\mathcal{T}}\sup_{d\left(\mathfrak{h},\mathfrak{h}^0\right)\leqslant\nu_{NT}}\left|\tilde{m}_{iT}\left(\tau,\mathfrak{h}\right)\right|\\
&=o_p\left(1\right)
\end{aligned}$$

The last line follows the Lemma S 4 with $\epsilon=0$ and Assumption S 5.1. □

Lemma S 6. *Under Assumption 2, Assumption S 2, and S 4,* $\mathfrak{M}\left(\tau,\mathfrak{h}\right)$ *is uniformly continuous, with respect to* $d\left(\cdot,\cdot\right)$*, at* $\mathfrak{h}^0$ *over* $\tau\in\mathcal{T}$.

Proof. From Lemma S 2's argument, $\left|\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}^\circ\right)\right]\right|\leqslant C^Jd\left(\mathfrak{h},\mathfrak{h}^\circ\right)$, for $\mathfrak{h}\in\mathfrak{J}$.
Let $\epsilon>0$. By the triangle inequality and Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left|\mathfrak{M}\left(\tau,\mathfrak{h}\right)-\mathfrak{M}\left(\tau,\mathfrak{h}^0\right)\right|\leqslant\frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}^\circ\right)\right]\right|^2\left\|W\right\|\\
&+4\frac{1}{N}\sum_{i=1}^{N}\left\|W\right\|LM^{\mathfrak{m}}\left|\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}\right)\right]-\mathbb{E}\left[\mathfrak{m}\left(w_{it},\tau,\mathfrak{h}^\circ\right)\right]\right|\\
&\leqslant4\left\|W\right\|LM^{\mathfrak{m}}C^Jd\left(\mathfrak{h},\mathfrak{h}^0\right)+\left\|W\right\|\left(C^J\right)^2\left(d\left(\mathfrak{h},\mathfrak{h}^0\right)\right)^2\\
&<\epsilon,
\end{aligned}$$

whenever $d\left(\mathfrak{h},\mathfrak{h}^0\right)<\delta$ with $\delta=\min\left\{\frac{\epsilon}{16\left\|W\right\|LM^{\mathfrak{m}}C^J},\frac{\sqrt{\epsilon}}{\sqrt{4\left\|W\right\|C^J}},\delta_J\right\}$. The uniformity comes from δ is independent of τ. □

Proof. (Proof of Theorem S 1) To apply Chen, Linton, and Keilegom (2003)'s Theorem 1, the proof verifies their five conditions. Their first condition is satisfied by the definition of $\hat{\tau}$ being the minimizer of $\hat{\mathfrak{M}}_{NT}\left(\tau,\hat{\mathfrak{h}}\right)$ over the parameter space of $\mathcal{T}$. Their second condition is verified

by Lemma S 1. Their third condition is verified by Lemma S 6. Their fourth condition is assumed by Assumption S 5.2. Finally, their fifth condition is verified by Lemma S 5. □

Section 3.2 Outline of the proof

1. Define

$$\tilde{r}(\tau)^L := \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) \underline{\omega}\left(z_{it}, \tau\right)\right).$$

   In absence of estimation error of $m(z_{it})$, $\tilde{r}_\tau^L$ is $\widehat{r_\tau^L}$. Hence, showing consistency of $\tilde{r}_\tau^L$ helps to derive the nonparametric estimation error of $\widehat{r_\tau^L}$. Subsequently, define $\tilde{R}(\omega, \tau) := b^L(\underline{\omega}(z_{it}, \tau))' \tilde{r}_\tau^L$.

2. The proof proceeds in three steps.

   (a) Show

$$(NT)^{\frac{1}{4}} \sup_{\tau \in \mathcal{T}} \sup_{\omega \in \mathcal{W}} \left| \tilde{R}(\omega, \tau) - R(\omega, \tau) \right| = o_p(1).$$

   (b) Show

$$(NT)^{\frac{1}{4}} \sup_{\tau \in \mathcal{T}} \sup_{\omega \in \mathcal{W}} \left| \hat{R}(\omega, \tau) - \tilde{R}(\omega, \tau) \right| = o_p(1),$$

   .

   (c) In conclusion, by triangle-inequality, Theorem R 1's uniform convergence holds.

Lemma R 1. *Under Assumption R 2, R 3, and R 4, there exist constant M_1, M_2, and M_3 such that,*

$$\left| b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) v_{it}^{\tau} - b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right) v_{it}^{\tau'} \right| \leqslant \sqrt{L} \xi_L^b \left(M_1 + M_2\right) \left|\tau' - \tau\right|$$

and

$$\left| b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' - b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right)' \right| \leqslant M_3 L \left(\xi_L^b\right)^2 \left|\tau - \tau'\right|.$$

Proof. Note that,

$$\frac{\partial}{\partial\tau}\left[b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}\right]=b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\frac{\partial}{\partial\tau}\left[\underline{\omega}\left(z_{it-1},\tau\right)-R\left(\underline{\omega}\left(z_{it-1},\tau\right),\tau\right)\right]$$
$$+\frac{db_l^L}{d\omega}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\frac{\partial\underline{\omega}}{\partial\tau}\left(z_{it-1},\tau\right)\left[\underline{\omega}\left(z_{it-1},\tau\right)-R\left(\underline{\omega}\left(z_{it-1},\tau\right),\tau\right)\right].$$
$$\frac{\partial}{\partial\tau}\left[b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b_s^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\right]=\frac{db_l^L}{d\omega}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\frac{\partial\underline{\omega}}{\partial\tau}\left(z_{it-1},\tau\right)b_s^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)$$
$$+\frac{db_s^L}{d\omega}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\frac{\partial\underline{\omega}}{\partial\tau}\left(z_{it-1},\tau\right)b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)$$

From Assumption R 2 and Assumption R 3, by Extreme Value Theorem, there exist constants M_1, M_2, and M_3 such that

1. $\sup_{\left(z_{it-1},\tau\right)\in\mathcal{Z}\times\mathcal{T}}\left|\frac{\partial}{\partial\tau}\left[\underline{\omega}\left(z_{it-1},\tau\right)-R\left(\underline{\omega}\left(z_{it-1},\tau\right),\tau\right)\right]\right|\leqslant M_1.$
2. $\sup_{\left(z_{it-1},\tau\right)\in\mathcal{Z}\times\mathcal{T}}\left|\frac{\partial\underline{\omega}}{\partial\tau}\left(z_{it-1},\tau\right)\left(\underline{\omega}\left(z_{it-1},\tau\right)-R\left(\underline{\omega}\left(z_{it-1},\tau\right),\tau\right)\right)\right|\leqslant M_2.$
3. $\sup_{\left(z_{it-1},\tau\right)\in\mathcal{Z}\times\mathcal{T}}\left|\frac{\partial\underline{\omega}}{\partial\tau}\left(z_{it-1},\tau\right)\right|\leqslant\frac{M_3}{2}.$

Then by applying the mean value theorem,

$$\begin{aligned}\left|b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}-b^L\left(\underline{\omega}\left(z_{it-1},\tau'\right)\right)v_{it}^{\tau'}\right|&\leqslant\sqrt{\sum_{l=1}^{L}\left(b_l^L\left(\underline{\omega}\left(z_{it-1},\overline{\tau}\right)\right)\right)^2M_1^2\left|\tau'-\tau\right|^2}\\&+\sqrt{\sum_{l=1}^{L}\left(\frac{\partial b_l^L}{\partial\omega}\left(\underline{\omega}\left(z_{it-1},\overline{\tau}\right)\right)\right)^2M_2^2\left|\tau'-\tau\right|^2}\\&\leqslant M_1\sqrt{L}\xi_L^b\left|\tau-\tau'\right|+M_2\sqrt{L}\xi_L^b\left|\tau-\tau'\right|\\\left|b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'-b^L\left(\underline{\omega}\left(z_{it-1},\tau'\right)\right)b^L\left(\underline{\omega}\left(z_{it-1},\tau'\right)\right)'\right|&\leqslant\sqrt{\sum_{s=1}^{L}\sum_{=1}^{L}\left(\left(\xi_L^b\right)^2M_3\left|\tau-\tau'\right|\right)^2}\\&=L\left(\xi_L^b\right)^2M_3\left|\tau-\tau'\right|\end{aligned}$$

□

Lemma R 2. *Under Assumption R 1, R 2, R 3, and R 4,*

$$\mathbb{P}\left(\sqrt{L}\xi_L^b\left(NT\right)^{\frac{1}{4}+r^{\psi}}\sup_{\tau\in\mathcal{T}}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}\right|>\epsilon\right)=o\left(1\right),$$

when $\frac{\mathfrak{v}_{N,T,L}\sqrt{L}\left(\xi_L^b\right)^2\left(NT\right)^{\frac{1}{4}+r^{\psi}}}{N}\to0$ *and* $\frac{\mathfrak{v}_{N,T,L}L\left(\xi_L^b\right)^4\left(NT\right)^{\frac{1}{2}+2r^{\psi}}}{NT}\to0$ *as* $N,T,L\to\infty$, *where* $\mathfrak{v}_{N,T,L}:=d_3\log\left(L\left(\xi_L^b\right)^2\left(NT\right)^{\frac{1}{4}+r^{\psi}}\right).$

Proof. First, the proof shows the convergence happens pointwise for $\tau\in\mathcal{T}$. Then the uniform convergence is derived from a covering argument. Let $\epsilon>0$.

Take $M=\max\{M_1,M_2\}$. Now set $\delta\left(\epsilon\right)=\frac{1}{4}\frac{\epsilon}{LM\left(\xi_L^b\right)^2\left(NT\right)^{\frac{1}{4}+r^{\psi}}}$. As $\mathcal{T}$ is compact, it admits a finite subcover of $\{B\left(\tau,\delta\left(\epsilon\right)\right)\}_{\tau\in\mathbb{s}}d_3$ The proof denotes the subcover as $\left\{B\left(\tau_j,\delta\left(\epsilon\right)\right)\right\}_{j=1}^{k(\epsilon)}$. As in Ai and Chen (2003), the covering number $k\left(\epsilon,N,T,L\right)\lessapprox\delta\left(\epsilon\right)^{-d_3}\lessapprox\left(L\left(\xi_L^b\right)^2\left(NT\right)^{\frac{1}{4}+r^{\psi}}\right)^{d_3}$. Re-label each subcover $B\left(\tau_j,\delta\left(\epsilon\right)\right)$ as $\mathfrak{C}_{\tau_j^{*}}$ where $\tau_j^{*}\in B\left(\tau_j,\delta\left(\epsilon\right)\right)\bigcap\mathcal{T}$.

For any $\tau \in \mathcal{T}$, there exists a $\tau_j^* \in \mathcal{T}$ such that $\tau \in \mathfrak{C}_{\tau_j^*}$, a cover in the finite subcover. By Lemma R 1,

$$\sqrt{L}\xi_L^b \left(NT\right)^{\frac{1}{4}+r^\psi} \left| b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} - b^L\left(\underline{\omega}\left(z_{it-1},\tau_j^*\right)\right) v_{it}^{\tau_j^*} \right| < \epsilon.$$

Therefore,

$$\begin{aligned}
&\left(NT\right)^{\frac{1}{4}+r^\psi} \sqrt{L}\xi_L^b \sup_{\tau\in\mathcal{T}} \left| \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=2}^{T} b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} \right| \\
&< \epsilon + \sqrt{L}\xi_L^b \left(NT\right)^{\frac{1}{4}+r^\psi} \sum_{j=1}^{k(\epsilon,N,T,L)} \left| \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=2}^{T} b^L\left(\underline{\omega}\left(z_{it-1},\tau_j^*\right)\right) v_{it}^{\tau_j^*} \right| \\
&\leqslant \epsilon + \sqrt{L}\xi_L^b \left(NT\right)^{\frac{1}{4}+r^\psi} \sum_{j=1}^{k(\epsilon,N,T,L)} \frac{1}{N}\sum_{i=1}^{N} \left| \frac{1}{T}\sum_{t=2}^{T} b^L\left(\underline{\omega}\left(z_{it-1},\tau_j^*\right)\right) v_{it}^{\tau_j^*} \right| \\
&\leqslant \epsilon + \sqrt{L}\xi_L^b \left(NT\right)^{\frac{1}{4}+r^\psi} \sum_{j=1}^{k(\epsilon,N,T,L)} \sum_{l=1}^{L} \frac{1}{N}\sum_{i=1}^{N} \left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau_j^*\right)\right) v_{it}^{\tau_j^*} \right|
\end{aligned}$$

From Assumption R 2 and R 3, $b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau}$ is bounded by $\xi_L^b C^v$ for some constant C^v. And,

$$\begin{aligned}
&\mathbb{E}\left[\left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} \right|^2 \right] = \frac{1}{T^2}\sum_{t=2}^{T}\sum_{s=2}^{T} \mathbb{E}\left[v_{it}^{\tau} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) b_l^L\left(\underline{\omega}\left(z_{is-1},\tau\right)\right) v_{is}^{\tau} \right] \\
&= \frac{1}{T^2}\sum_{t=2}^{T}\sum_{s=2}^{T} Cov\left(v_{it}^{\tau} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right), b_l^L\left(\underline{\omega}\left(z_{is-1},\tau\right)\right) v_{is}^{\tau} \right) \\
&\leqslant \frac{\left(C^v\right)^2 \left(\xi_L^b\right)^2}{T^2} \sum_{t=2}^{T}\sum_{s=2}^{T} 12\left(\rho_i^{z,\omega}\left(\alpha,|t-s|\right)\right)^{\frac{1}{3}} \\
&\leqslant 12\frac{\left(\xi_L^b\right)^2 \left(C^v\right)^2 C^f}{T}
\end{aligned}$$

The first equality comes from the conditional independence assumption. The second last line uses the Davydov inequality and the iterated laws of expectation. And the last line uses Assumption R 1.2. Thus

$$\mathbb{E}\left[\left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} \right|^2 \right] = O\left(\frac{\left(\xi_L^b\right)^2}{T} \right).$$

Then Bernstein inequality of independent processes provides

$$\mathbb{P}\left(\sqrt{L}\xi_L^b \left(NT\right)^{\frac{1}{4}+r^\psi} \frac{1}{N}\sum_{i=1}^{N} \left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau_j^*\right)\right) v_{it}^{\tau_j^*} \right| > \epsilon' \middle| \alpha \right) \leqslant 2\exp\left(-\frac{\left(\epsilon'\right)^2}{\kappa_d} \right),$$

where

$$\begin{aligned}
\kappa_d\left(\alpha\right) \leqslant\ & \frac{2L\left(NT\right)^{\frac{1}{2}+2r^\psi}\left(\xi_L^b\right)^2}{N} \sup_{i\in\{1,\ldots,N\}} \left[\mathbb{E}\left[\left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} \right|^2 \middle| \alpha \right] - \left(\mathbb{E}\left[\left| \frac{1}{T}\sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right) v_{it}^{\tau} \right| \middle| \alpha \right] \right)^2 \right] \\
&+ \frac{2}{3} C^v \frac{\sqrt{L}\left(\xi_L^b\right)^2 \left(NT\right)^{\frac{1}{4}+r^\psi} \epsilon'}{N}.
\end{aligned}$$

By the iterated laws of expectation, $\kappa_d(\alpha) \leqslant O_p\left(\frac{L\left(\xi_L^b\right)^4 (NT)^{\frac{1}{2}+2r^\psi}}{NT}\right) + O_p\left(\frac{\sqrt{L}\left(\xi_L^b\right)^2 (NT)^{\frac{1}{4}+r^\psi}}{N}\right)$.

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{L}\xi_L^b (NT)^{\frac{1}{4}+r^\psi} \sum_{j=1}^{k(\epsilon,N,T,L)} \sum_{l=1}^{L} \frac{1}{N} \sum_{i=1}^{N} \left|\frac{1}{T} \sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1}, \tau_j^*\right)\right) v_{it}^{\tau_j^*}\right| > \epsilon'\right) \\
&\leqslant \sum_{j=1}^{k(\epsilon,N,T,L)} \sum_{l=1}^{L} \mathbb{E}\left[\mathbb{P}\left(\sqrt{L}\xi_L^b (NT)^{\frac{1}{4}+r^\psi} \frac{1}{N} \sum_{i=1}^{N} \left|\frac{1}{T} \sum_{t=2}^{T} b_l^L\left(\underline{\omega}\left(z_{it-1}, \tau_j^*\right)\right) v_{it}^{\tau_j^*}\right| > \epsilon' \middle| \alpha\right)\right] \\
&\leqslant \mathbb{E}\left[\exp\left(\log\left(k\left(\epsilon, N, T, L\right)\right) + \log(L) - \frac{\left(\epsilon'\right)^2}{O_p\left(\frac{L\left(\xi_L^b\right)^4 (NT)^{\frac{1}{4}+r^\psi}}{NT}\right) + O_p\left(\frac{\sqrt{L}\left(\xi_L^b\right)^2 (NT)^{\frac{1}{4}+r^\psi}}{N}\right)}\right)\right]
\end{aligned}$$

By the continuous mapping theorem, the last line is $o(1)$ when $\frac{\sqrt{L}\mathfrak{d}_{N,T,L}\left(\xi_L^b\right)^2 (NT)^{\frac{1}{4}+r^\psi}}{N} \to 0$ and $\frac{L\mathfrak{d}_{N,T,L}\left(\xi_L^b\right)^4 (NT)^{\frac{1}{4}+r^\psi}}{NT} \to 0$ as $N, T, L \to \infty$. These are provided under Assumption R 6. □

Lemma R 3. *Under Assumption R 1, R 2, R 3, R 4, and R 5,*

$$\sup_{\tau\in\mathcal{T}} (NT)^{r^\psi} \left|\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} \mathfrak{b}^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' - \psi^{bb,L}(\alpha, \tau, T)\right| = o(1),$$

when $\frac{\mathfrak{d}_{N,T,L}(NT)^{2r^\psi}\left(\xi_L^b\right)^4}{NT} \to 0$, $\frac{\mathfrak{d}_{N,T,L}(NT)^{r^\psi}\left(\xi_L^b\right)^4}{N} \to 0$, *and* $\frac{T}{N} \to 0$, *where* $\mathfrak{d}_{N,T,L} = d_3 \log\left(L\left(\xi_L^b\right)^2 (NT)^{r^\psi}\right)$.

Proof. Let $\mathfrak{b}^L\left(z_{it-1}, \tau\right) := b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' - \mathbb{E}\left[b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) | \alpha\right]$.

From triangle inequality and Lemma R 1,

$$\begin{aligned}
&\left|\mathfrak{b}^L\left(z_{it-1}, \tau\right) - \mathfrak{b}^L\left(z_{it-1}, \tau'\right)\right| \leqslant \left|b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' - b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right)'\right| \\
&+ \mathbb{E}\left[\left|b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau\right)\right)' - b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right) b^L\left(\underline{\omega}\left(z_{it-1}, \tau'\right)\right)'\right| \middle| \alpha\right] \leqslant 2M_3 L\left(\xi_L^b\right)^2 \left|\tau - \tau'\right|.
\end{aligned}$$

Let $\epsilon > 0$ and set $\delta(\epsilon) = \frac{1}{4} \frac{\epsilon}{2M_3 L\left(\xi_L^b\right)^2 (NT)^{r^\psi}}$. As in Lemma R 2, find the $k(\epsilon, N, T, L)$ covers $\mathfrak{C}_{\tau_1^*}, \ldots, \mathfrak{C}_{\tau_{k(\epsilon)}^*}$ for $\mathcal{T}$ such that if $\tau \in \mathcal{T}$ then $\left|\tau - \tau_j^*\right| < \delta(\epsilon)$ for some $\tau_j^* \in \mathcal{T}$, $j = 1, \ldots, k(\epsilon)$.

As in Lemma R 2's argument,

$$\sup_{\tau\in\mathcal{T}} \frac{(NT)^{r^\psi}}{NT} \left|\sum_{i=1}^{N} \sum_{t=2}^{T} \mathfrak{b}^L\left(z_{it-1}, \tau\right)\right| < \epsilon + (NT)^{r^\psi} \sum_{j=1}^{k(\epsilon)} \sum_{l=1}^{L} \sum_{v=1}^{L} \sum_{i=1}^{N} \frac{1}{N} \left|\frac{1}{T} \sum_{t=2}^{T} \mathfrak{b}_{lv}^L\left(z_{it-1}, \tau_j^*\right)\right|.$$

Here, $k(\epsilon) \lesssim \left(L\left(\xi_L^b\right)^2 (NT)^{r^\psi}\right)^{d_3}$.

A similar argument in Lemma 11's proof provides $\mathbb{E}\left[\frac{1}{T} \sum_{t=2}^{T} \left|\mathfrak{b}_{lv}^L\left(z_{it-1}, \tau_j^*\right)\right|^2\right] = O\left(\frac{\left(\xi_L^b\right)^4}{T}\right)$.

Then an argument similar to Lemma R 2's proof provides

$$\mathbb{P}\left((NT)^{r^\psi} \sum_{j=1}^{k(\epsilon)} \sum_{l=1}^{L} \sum_{v=1}^{L} \sum_{i=1}^{N} \frac{1}{N} \left|\frac{1}{T} \sum_{t=2}^{T} \mathfrak{b}_{lv}^L\left(z_{it-1}, \tau_j^*\right)\right| > \epsilon'\right) = o(1)$$

when $\frac{\mathfrak{d}_{N,T,L}(NT)^{2r^\psi}\left(\xi_L^b\right)^4}{NT} \to 0$ and $\frac{\mathfrak{d}_{N,T,L}(NT)^{r^\psi}\left(\xi_L^b\right)^4}{N} \to 0$. As ϵ' is arbitrary, $\sup_{\tau\in\mathcal{T}} \frac{(NT)^{r^\psi}}{NT} \left|\sum_{i=1}^{N} \sum_{t=2}^{T} \mathfrak{b}^L\left(z_{it-1}, \tau\right)\right| = o_p(1)$.

From triangle inequality,

$$\sup_{\tau\in\mathcal{T}}(NT)^{r^{\psi}}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'-\psi^{bb,L}\left(\alpha,\tau,T\right)\right|\leqslant\sup_{\tau\in\mathcal{T}}\frac{(NT)^{r^{\psi}}}{NT}\left|\sum_{i=1}^{N}\sum_{t=2}^{T}{}_{b}{}^{L}\left(z_{it-1},\tau\right)\right|$$
$$+\sup_{\tau\in\mathcal{T}}(NT)^{r^{\psi}}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbb{E}\left[b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'|\alpha\right]-\psi^{bb,L}\left(\alpha,\tau,T\right)\right|$$

From Assumption R 5.1, the last term is $O\left(\frac{(NT)^{r^{\psi}}}{\sqrt{N}}\right)\to 0$, as $\frac{T}{N}\to 0$. This completes the proof. □

Lemma R 4. *Under Assumption R 1, R 2, R 3, R 4, R 5, and R 6,*

$$\sup_{\tau\in\mathcal{T}}\frac{1}{(NT)^{r^{\psi}}}\left|\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'\right]^{-1}\right|=O_{p}\left(1\right).$$

Proof. By Assumption R 5.1, $\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'\right]^{-1}$ exists provided N is large enough. For simplicity, the proof assumes the inverse exists.

Using triangle inequality and then re-arranging,

$$\sup_{\tau\in\mathcal{T}}\frac{1}{(NT)^{r^{\psi}}}\left|\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'\right]^{-1}\right|$$
$$\leqslant\frac{\sup_{\tau\in\mathcal{T}}\left|\left[(NT)^{r^{\psi}}\ \psi^{bb,L}\left(\alpha,\tau,T\right)\right]^{-1}\right|}{1-\sup_{\tau\in\mathcal{T}}(NT)^{r^{\psi}}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'-\psi^{bb,L}\left(\alpha,\tau\right)\right|\sup_{\tau\in\mathcal{T}}\left|\left[(NT)^{r^{\psi}}\ \psi^{bb,L}\left(\alpha,\tau,T\right)\right]^{-1}\right|}.$$

From Lemma R 3 and Assumption R 5.2, the upper bound is $O_{p}\left(1\right)$.

□

Lemma R 5. *Under Assumption R1, R 2, R 3, R 4, R 5, and R 6,*

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}(NT)^{\frac{1}{4}}\left|b^{L}\left(\omega\right)'\left(\hat{r}\left(\tau\right)-r^{0,L}\left(\tau\right)\right)\right|=o_{p}\left(1\right).$$

Proof. Let $\mathfrak{u}:=\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'\right]^{-1}$.

By applying Assumption R 4.1 and Lemma R 4,

$$\sup_{\tau\in\mathcal{T}}\sup_{w\in\mathcal{W}}\left|(NT)^{\frac{1}{4}}\ b^{L}\left(\omega\right)\mathfrak{u}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\left(R\left(w,\tau\right)-b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'r^{0,L}\left(\tau\right)\right)\right)\right|$$
$$=O_{p}\left(\frac{(NT)^{\frac{1}{4}+r^{\psi}}\ L\left(\xi_{L}^{b}\right)^{2}}{L^{\mu}R}\right).$$

The last line is $o_{p}\left(1\right)$ because of Assumption R 6's $\frac{(NT)^{\frac{1}{4}+r^{\psi}}\left(\xi_{L}^{b}\right)^{2}}{L^{\mu}R}\to 0$. Then from Lemma R 4,

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|(NT)^{\frac{1}{4}}\ b^{L}\left(\omega\right)\mathfrak{u}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}\right)\right|\leqslant O_{p}\left(1\right)\sup_{\tau\in\mathcal{T}}\left|\frac{\xi_{L}^{b}\left(NT\right)^{\frac{1}{4}+r^{\psi}}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}b^{L}\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}\right|$$
$$=o_{p}\left(1\right)$$

The last line comes from Lemma R 2 and Assumption R 6.

Since,

$$b^L(\omega)'\left(\bar{r}(\tau)-r^{0,L}(\tau)\right)=b^L(\omega)'\mathfrak{u}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)\left(R(w,\tau)-b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'r^{0,L}(\tau)\right)\right)$$
$$+b^L(\omega)'\mathfrak{u}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)v_{it}^{\tau}\right)$$

Therefore, $\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|b^L(\omega)'\left(\bar{r}(\tau)-r^{0,L}(\tau)\right)\right|=o_p\left((NT)^{-\frac{1}{4}}\right)$. □

Lemma R 6. *Under Assumption R 1, R 2, R 3, R 4, R 5, and R 6,*

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}(NT)^{\frac{1}{4}}\left|\tilde{R}(\omega,\tau)-R(\omega,\tau)\right|=o_p(1),$$

as $N,T,L\to\infty$.

Proof. By triangle inequality,

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}(NT)^{\frac{1}{4}}\left|\tilde{R}(\omega,\tau)-R(\omega,\tau)\right|\leqslant(NT)^{\frac{1}{4}}\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|b^L(\omega)'\left(r^{0,L}(\tau)-\bar{r}(\tau)\right)\right|$$
$$+(NT)^{\frac{1}{4}}\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|b^L(\omega)'r^{0,L}(\omega)-R(\omega,\tau)\right|$$

The RHS's first term is $o_p(1)$ by Lemma R 5 and the RHS's second term is $o_p(1)$ from Assumption R 6. □

As defined in Assumption R 3's discussion, $\left|\hat{\underline{\omega}}(z_{it},\tau)-\underline{\omega}(z_{it},\tau)\right|<\delta$ implies $\hat{\underline{\omega}}(z_{it},\tau)\in\mathcal{W}$. Since $\left|\hat{\underline{\omega}}(z_{it},\tau)-\underline{\omega}(z_{it},\tau)\right|=\left|\widehat{m}(z_{it})-m(z_{it})\right|$, $\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|<\delta$ implies all $\hat{\underline{\omega}}(z_{it},\tau)\in\mathcal{W}$. For Lemma R 7, $\hat{\underline{\omega}}(z_{it},\tau)\in\mathcal{W}$ is assumed.

Lemma R 7. *Conditional on* $\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|<\delta$, *under Assumption R 1, R 2, R 3, and R 6,*

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}(NT)^{\frac{1}{4}}\left|\widehat{R}(\omega,\tau)-\tilde{R}(\omega,\tau)\right|=o_p(1).$$

Proof.

$$\sup_{\tau\in\mathcal{T}}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)b^L\left(\underline{\omega}\left(z_{it-1},\tau\right)\right)'-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}b^L\left(\hat{\underline{\omega}}\left(z_{it-1},\tau\right)\right)b^L\left(\hat{\underline{\omega}}\left(z_{it-1},\tau\right)\right)'\right|$$
$$\leqslant\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L(\underline{\omega}(z,\tau))b^L(\underline{\omega}(z,\tau))'-b^L(\hat{\underline{\omega}}(z,\tau))b^L(\hat{\underline{\omega}}(z,\tau))'\right|$$
$$\leqslant\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L(\underline{\omega}(z,\tau))\right|\left|b^L(\underline{\omega}(z,\tau))'-b^L(\hat{\underline{\omega}}(z,\tau))\right|$$
$$+\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L(\hat{\underline{\omega}}(z,\tau))\right|\left|b^L(\underline{\omega}(z,\tau))'-b^L(\hat{\underline{\omega}}(z,\tau))\right|$$
$$\leqslant O_p\left(L\left(\xi_L^b\right)^2\Delta\right)$$

The second inequality comes from triangle inequality and the final line comes from the Taylor's expansion.

Thus, under $L\left(\xi_L^b\right)^2\Delta \to 0$, doing a similar expansion as above provides

$$\begin{aligned}
\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|\widehat{R}(\omega,\tau)-\tilde{R}(\omega,\tau)\right| &\leqslant O_p(1)\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L\left(\underline{\widehat{\omega}}(z,\tau)\right)\underline{\widehat{\omega}}(z,\tau)-b^L\left(\underline{\omega}(z,\tau)\right)\underline{\omega}(z,\tau)\right| \\
&\leqslant O_p(1)\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L\left(\underline{\widehat{\omega}}(z,\tau)\right)\right|\left\|\underline{\widehat{\omega}}(z,\tau)-\underline{\omega}(z,\tau)\right\| \\
&+ O_p(1)\sup_{g\in\{1,\ldots,G^0\}}\sup_{\tau\in\mathcal{T}}\sup_{z\in\mathcal{Z}}\left|b^L\left(\underline{\omega}(z,\tau)\right)-b^L\left(\underline{\widehat{\omega}}(z,\tau)\right)\right| \\
&= O_p\left(\sqrt{L}\xi_L^b\Delta\right)
\end{aligned}$$

The first inequality follows from the previous result and the estimators' formula. The second inequality uses the fact of $\mathcal{W}$ being compact and the triangle inequality. The final line uses the mean-value theorem.

Finally, Assumption R 6 provides $(NT)^{\frac{1}{4}}\sqrt{L}\xi_L^b\Delta \to 0$ and $L\left(\xi_L^b\right)^2\Delta \to 0$ to complete the proof. □

Proof of Theorem R 1. From the law of total expectation,

$$\begin{aligned}
&\mathbb{E}\left[(NT)^{\frac{1}{4}}\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|\widehat{R}(\omega,\tau)-\tilde{R}(\omega,\tau)\right|\right] \\
&= \mathbb{E}\left[(NT)^{\frac{1}{4}}\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|\widehat{R}(\omega,\tau)-\tilde{R}(\omega,\tau)\right|\middle|\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|<\delta\right]\mathbb{P}\left(\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|<\delta\right) \\
&+ \mathbb{E}\left[\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}\left|\widehat{R}(\omega,\tau)-\tilde{R}(\omega,\tau)\right|\middle|\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|>\delta\right](NT)^{\frac{1}{4}}\mathbb{P}\left(\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|>\delta\right)
\end{aligned}$$

From Assumption R 6 and Markov inequality, $(NT)^{\frac{1}{4}}\mathbb{P}\left(\sup_{z\in\mathcal{Z}}\left|\widehat{m}(z)-m(z)\right|>\delta\right)=o(1)$. Lemma R 6 implies the first term is $o(1)$. Then the result is immediate from Lemma R 6 and the triangle inequality. □

Appendix - Chapter 4

Proof of Lemma 1. Define the population criterion

$$\begin{aligned} Q_N(\theta, G, H) &= N^{-1} \sum_{i=1}^{N} Q_i(\theta, g_i, h_i), \text{ where} \\ Q_i(\theta, g_i, h_i) &= E[m(w_{it}; \alpha(g_i), \beta(h_i), \lambda)]' W_i E[m(w_{it}; \alpha(g_i), \beta(h_i), \lambda)]. \end{aligned} \tag{A.1}$$

By Assumption W and (4.14), we have the uniform convergence

$$\sup_{(\theta,G,H)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} |\widehat{Q}_N(\theta, G, H) - \widehat{Q}_N(\theta, G, H)| = o_p(1). \tag{A.2}$$

Define

$$\begin{aligned} d(\theta, G, H) &= N^{-1} \sum_{i=1}^{N} d_i(\theta_i), \text{ where} \\ d_i(\theta_i) &= \left(\alpha(g_i) - \alpha^0(g_i^0)\right)^2 + \left(\beta(h_i) - \beta^0(h_i^0)\right)^2 + ||\lambda - \lambda^0||^2. \end{aligned} \tag{A.3}$$

We show that, for any $\delta > 0$, there exists $\varepsilon > 0$ such that

$$\inf_{d(\theta,G,H)>\delta} Q_N(\theta, G, H) \geqslant \varepsilon. \tag{A.4}$$

Given that θ_i has a compact support Θ for all i, let $C = \sup_i \sup_{\theta_i\in\Theta} d_i(\theta_i) < \infty$. Let $S = \{i : d_i(\theta_i) > \delta/2\}$ and $N_S = \sum_{i=1}^{N} 1\{i \in S\}$. Note that $d_i(\theta_i) \leqslant C$ for $i \in S$ and $d_i(\theta_i) \leqslant \delta/2$ for $i \notin S$. Thus, $N_S C + (N - N_S)\delta/2 \geqslant N d(\theta, G, H) \geqslant N\delta$, which implies that $N_S \geqslant N\delta/(2C - \delta) > N\delta/(2C)$. Then,

$$\inf_{d(\theta,G,H)>\delta} Q_N(\theta, G, H) \geqslant \inf_{d(\theta,G,H)>\delta} N^{-1} \sum_{i\in S} Q_i(\theta, g_i, h_i) \geqslant \frac{N_s}{N} \min_{i\in S} Q_i(\theta, g_i, h_i) \geqslant \frac{\delta}{2C}\varepsilon^*, \tag{A.5}$$

where the last step holds because $\min_{i\in S} Q_i(\theta, g_i, h_i) \geqslant \varepsilon^*$ for some $\varepsilon^* > 0$ by Assumption ID

and W. Thus, the identification condition for $Q_N(\theta, G, H)$ in (A.4) holds with $\varepsilon = \delta\varepsilon^*/(2C)$. Results in (A.5) is analogous to Lemma A.4 in Liu et al. (2018).

Finally, we show the consistency result by combining (A.2) and (A.4). For any $\delta > 0$, there exists $\varepsilon > 0$, such that

$$\begin{aligned} P\left\{d(\hat{\theta}, \widehat{G}, \widehat{H}) > \delta\right\} &\leqslant P\left\{Q_N(\hat{\theta}, \widehat{G}, \widehat{H}) \geqslant \varepsilon\right\} \\ &= P\left\{d_1 + d_2 + d_3 \geqslant \varepsilon\right\}, \end{aligned} \tag{A.6}$$

where

$$\begin{aligned} d_1 &= Q_N(\hat{\theta}, \widehat{G}, \widehat{H}) - \widehat{Q}_N(\hat{\theta}, \widehat{G}, \widehat{H}), \\ d_2 &= \widehat{Q}_N(\hat{\theta}, \widehat{G}, \widehat{H}) - \widehat{Q}_N(\theta^0, G^0, H^0), \\ d_3 &= \widehat{Q}_N(\theta^0, G^0, H^0) - Q_N(\theta^0, G^0, H^0). \end{aligned} \tag{A.7}$$

Because $d_2 \leqslant 0$ by definition of the estimator and $d_1 = o_p(1)$ and $d_3 = o_p(1)$ by (A.2), (A.6) implies that $P\{d(\hat{\theta}, \widehat{G}, \widehat{H}) > \delta\} \to 0$ for any $\delta > 0$. This completes the proof. □

Proof of Lemma 2. Given Lemma 1 and Assumption S, this Lemma follows from the same arguments used to show Lemma B.3 of BM. The arguments can be applied to α and β separately in our set-up. There is no need to take sample average here because our parameters are not time-varying. Lemma B.3 also shows how to relabel the groups and shows that this is a one-to-one mapping with probability approaching 1. □

Proof of Theorem 5. Let $E_W = 1\{\max_i |W_{iNT} - W_i| \leqslant \eta\}$ for some small constant η, Assumption W shows that $E_W = 1$ with probability approaching 1. Conditional on $E_W = 1$, for $(g_i, h_i) \neq (g_i^0, h_i^0)$,

we have shown in (4.16)-(4.18) that

$$\begin{aligned} & P\left\{\hat{g}_i = g_i, \hat{h}_i = h_i\right\} \\ \leqslant\ & P\left\{\widehat{Q}_i(\hat{\theta}, g_i, h_i) < \widehat{Q}_i(\hat{\theta}, g_i^0, h_i^0)\right\} \\ \leqslant\ & P\left\{c_1 \left|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\hat{\theta}, g_i, h_i)\right|^2 \leqslant c_2 \left|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\hat{\theta}, g_i^0, h_i^0)\right|^2\right\} \end{aligned} \tag{A.8}$$

for constants $c_2 > c_1 > 0$. Using the decomposition in (4.19) and the triangle inequality,

$$\left|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\hat{\theta}, g_i, h_i)\right|^2 \geqslant \left|b_i(\hat{\theta}, g_i, h_i)\right| - \left|\delta_i(\hat{\theta}, g_i, h_i)\right|^2, \tag{A.9}$$

where

$$\begin{aligned} \delta_i(\hat{\theta}, g_i, h_i) &= \frac{1}{T}\sum_{t=1}^{T} m_{it}(\hat{\theta}, g_i, h_i) - E[m_{it}(\hat{\theta}, g_i, h_i)], \\ b_i(\hat{\theta}, g_i, h_i) &= E[m_{it}(\hat{\theta}, g_i, h_i)]. \end{aligned}$$

By a similarly decomposition,

$$\left|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\hat{\theta}, g_i^0, h_i^0)\right|^2 \leqslant \left|b_i(\hat{\theta}, g_i^0, h_i^0)\right| + \left|\delta_i(\hat{\theta}, g_i^0, h_i^0)\right|^2. \tag{A.10}$$

Below we analyze the four terms $\delta_i(\hat{\theta}, g_i, h_i),\ b_i(\hat{\theta}, g_i, h_i), \delta_i(\hat{\theta}, g_i^0, h_i^0), b_i(\hat{\theta}, g_i^0, h_i^0)$.

For $\hat{\theta} \in N_\eta = \{\theta \in \Theta : ||\theta - \theta_0||^2 \leqslant \eta^2\}$, we have

$$\begin{aligned} \left|b_i(\hat{\theta}, g_i, h_i)\right|^2 &= \left|E[m_{it}(\hat{\theta}, g_i, h_i)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right|^2 \\ &\geqslant b_{1,i}\left(\alpha_g^0, \beta_h^0\right) - b_{2,i}\left(\alpha_g, \beta_h\right) \end{aligned} \tag{A.11}$$

where

$$\begin{aligned} b_{1,i}\left(\theta^0, g_i, h_i\right) &= \left|E[m_{it}\left(\theta^0, g_i, h_i\right)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right|^2, \\ b_{2,i}\left(\hat{\theta}, g_i, h_i\right) &= \left|E[m_{it}(\hat{\theta}, g_i, h_i)] - E[m_{it}(\theta^0, g_i, h_i)]\right|^2, \end{aligned} \tag{A.12}$$

where the first term $b_{1,i}(\theta^0, g_i, h_i)$ is due to misspecification of group and the second term $b_{2,i}(\hat{\theta}, g_i, h_i)$ is due to the estimation error between $\hat{\theta}$ and θ^0. By Assumption ID and S,

$$b_{1,i}\left(\theta^0, g_i, h_i\right) \geqslant m_0 \tag{A.13}$$

for some $m_0 > 0$ for any $(g_i, h_i) \neq (g_i^0, h_i^0)$. By Assumption R(iii),

$$b_{2,i}\left(\hat{\theta}, g_i, h_i\right) \leqslant M_0 \eta^2 \tag{A.14}$$

for some $M_0 < \infty$. Therefore,

$$\left|b_i(\hat{\theta}, g_i, h_i)\right|^2 \geqslant m_0 - M_0 \eta^2. \tag{A.15}$$

Similarly, we have

$$\begin{aligned} \left|b_i(\hat{\theta}, g_i^0, h_i^0)\right| &= \left|E[m_{it}\left(\hat{\theta}, g_i^0, h_i^0\right)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right|^2 \\ &\leqslant M_0 \eta^2. \end{aligned} \tag{A.16}$$

Combining (A.8) with (A.9), (A.10), (A.15), (A.16), we obtain

$$\begin{aligned} & P_{i,gh}(\hat{\theta}) \\ \leqslant\ & P\left\{c_1 m_0 - c_1 M_0 \eta^2 - c_2 M_0 \eta^2 \leqslant c_1 \left|\delta_i(\hat{\theta}, g_i, h_i)\right|^2 + c_2 \left|\delta_i(\hat{\theta}, g_i^0, h_i^0)\right|^2\right\}. \end{aligned} \tag{A.17}$$

Take $\eta > 0$ small enough such that

$$s = c_1 m_0 - c_1 M_0 \eta^2 - c_2 M_0 \eta^2 > 0. \tag{A.18}$$

Note that $\delta_i(\hat{\theta}, g_i, h_i)$ and $\delta_i(\hat{\theta}, g_i^0, h_i^0)$ both are differences between sample mean and population mean. Under Assumption R,

$$\begin{aligned} \max_{1 \leqslant i \leqslant N} P\left\{c_1 \left|\delta_i(\hat{\theta}, g_i, h_i)\right|^2 \geqslant s\right\} &= o(N^{-1}), \\ \max_{1 \leqslant i \leqslant N} P\left\{c_2 \left|\delta_i\left(\hat{\theta}, g_i, h_i\right)\right|^2 \geqslant s\right\} &= o(N^{-1}), \end{aligned} \tag{A.19}$$

by Lemma S1.2(ii) of SSP. Therefore, for any $(g_i, h_i) \neq (g_i^0, h_i^0)$,

$$\max_{1 \leqslant i \leqslant N} P\left\{\hat{g}_i = g_i, \hat{h}_i = h_i\right\} = o(N^{-1}) \tag{A.20}$$

for $\hat{\theta} \in N_\eta$. Because g_i and h_i both have finite support, we obtain

$$\max_{1 \leqslant i \leqslant N} P\left\{\hat{g}_i \neq g_i^0, \hat{h}_i \neq h_i^0\right\} = o(N^{-1}) \tag{A.21}$$

for $\hat{\theta} \in N_\eta$.

Finally, conditional on $\hat{\theta} \in N_\eta$ and $E_W = 1$, we have

$$\begin{aligned} & P\left\{\widehat{G} = G^0 \text{ and } \widehat{H} = H^0\right\} \\ = \; & 1 - P\left\{1\left\{(\hat{g}_i, \hat{h}_i) \neq (g_i^0, h_i^0)\right\} \text{ for some } i\right\} \\ \geqslant \; & 1 - N \max_{1 \leqslant i \leqslant N} P\left\{(\hat{g}_i, \hat{h}_i) \neq (g_i^0, h_i^0)\right\} \\ \to \; & 1. \end{aligned} \tag{A.22}$$

By Lemma 2 and Assumption W, $P\{\hat{\theta} \in N_\eta\} \to 1$ and $P\{E_W = 1\} \to 1$, which gives the desirable

result together with (A.22). ▫

Proof of Theorem 6. Because $\widehat{G} = G_0$ and $\widehat{H} = H_0$ with probability approaching 1, $\tilde{\theta}$ has the same asymptotic distribution as the oracle estimator $\overline{\theta}$ that is obtained by assuming G_0 and H_0 are known, i.e.,

$$\overline{\theta} = \arg\min_{\theta \in \overline{\Theta}} \overline{Q}(\theta), \text{ where } \overline{Q}(\theta) = \overline{m}(\theta)' W_{NT} \overline{m}(\theta), \tag{A.23}$$

with

$$\overline{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i^0), \beta(h_i^0), \lambda\right). \tag{A.24}$$

Now we derive the asymptotic distribution of $\overline{\theta}$. This is a standard GMM problem. By Assumption ID, E(ii), and (4.14), we have the typical identification and uniform convergence conditions for the consistency of $\overline{\theta}$. To get the asymptotic distribution, it is sufficient to show for some $\eta > 0$,

$$N^{-1} \sum_{i=1}^{N} \sup_{||\theta_i - \theta_i^0|| \leqslant \eta} \left| T^{-1} \sum_{t=1}^{T} m_\theta\left(w_{it}; \theta_i\right) - E\left[m_\theta\left(w_{it}; \theta_i\right)\right] \right| \to_p 0 \tag{A.25}$$

and

$$(NT)^{-1/2} \sum_{i=1}^{N} \sum_{t=1}^{T} m\left(w_{it}; \theta_i^0\right) \to_d N(0, \Omega) \tag{A.26}$$

as $N, T \to \infty$. The first result in (A.25) follows from a uniform convergence over i, which is obtained by applying Lemma S1.2(iii) of SSP under Assumption R and E(iii). The second result in (A.26) follows from verifying a Lindeberg-Feller central limit theorem. Lemma S1.12 of SSP proves a result of the same form and provide the details of the verification, see p.29 of the Supplement to SSP. This completes the proof. ▫

Verification of Assumptions for the Production Function Example.

We first verify Assumption ID. For any $\theta_i = (a_i, b_i, c_i, \rho)$, we have

$$\Delta y_{it}(\rho) = a_i^0(1-\rho) + b_i^0 v_{it}(\rho) + c_i^0 k_{it}(\rho) + (\rho_0 - \rho)\,\omega_{it-1} + \xi_{it} + \varepsilon_{it} - \rho\varepsilon_{it-1}, \tag{A.27}$$

and

$$\begin{aligned} & E\left[z_{it}\left(\Delta y_{it}(\rho) - a_i(1-\rho) - b_i \Delta v_{it}(\rho) - c_i \Delta k_{it}(\rho)\right)\right] \\ = \;& E\left[z_{it}\left(\left(a_i^0 - a_i\right)(1-\rho) + \left(b_i^0 - b_i\right)\Delta v_{it}(\rho) + \left(c_i^0 - c_i\right)\Delta k_{it}(\rho) + (\rho_0 - \rho)\,\omega_{it-1}\right)\right] \end{aligned} \tag{A.28}$$

under condition (i). Assumption ID holds under $\mu_{\min}(E[z_{it}x_{it}(\rho)]') \geqslant \delta > 0$ and $\rho < 1$. Assumption R(i)-R(ii) holds automatically under conditions (i) and (ii). The first order derivative is

$$m_\theta(w_{it}; \theta_i) = -z_{it}[(1-\rho), \Delta v_{it}(\rho), \Delta k_{it}(\rho), y_{it-1} - a_i - b_i v_{it-1} - c k_{it-1}]. \tag{A.29}$$

Assumption R(iii) and E(iii) holds under $E||z_{it}d_{it}||^q \leqslant C$. □

Appendix - Misc

Monte Carlo

Monte Carlo: data-generating process

Here, I fully detail the data-generating process for the Monte Carlo simulation. The model is an extension of Ackerberg, Caves, and Frazer (2015)'s Monte Carlo DGP1 simulation. I use their parametric choices unless I specify otherwise. They chose the parameter values to match across-firms variation accounting for 95% of capital's variation and labor predicting 50% of capital's variation to match the stylized moments from the Levinsohn and Petrin (2003)'s Chilean data set.

In the simulated sector, N firms maximize expected profit by choosing variables M_{it}, L_{it} and K_{it}. Conditional on productivity, wage, capital-adjustment cost, and capital stock, each firm faces the same infinite horizon problem with a discount rate 0.95. It is a price-taker and uses the Leontiff production technology,

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^{w}_{g_i t}\right) \min\{M_{it} + M_{it}^2, \exp\left(\eta_{it}\right) K_{it}^{0.4} L_{it}^{0.6}\}.$$

The output price and intermediate material input price are time-invariant. Also, their difference is normalized to one.

The firm faces no dynamic constraints in choosing labour and intermediate material inputs. And it pays its workers at the wage W_{it} with the transition dynamic $\log(W_{it}) = 0.3 \log(W_{it-1}) + \xi^{W}_{it}$, where $\xi^{W}_{it} \overset{iid}{\sim} N(0, 0.0091)$.

The firm uses owned capital (there is no rent payment in its profit function) and invests I_{it} new capital at the end of period t. Newly invested capital is only effective in the next period, i.e. $I_{it} := K_{it+1} - 0.8 K_{it}$. This investment comes at a cost $\frac{\phi_i}{2} I_{it}^2$, where $\frac{1}{\phi_i} \overset{iid}{\sim} \log(N(0, 0.36))$.

The productivity process η_{it} is an AR(1) processes, with parametric specification $\eta_{it} = 0.7\eta_{it-1} + \xi_{it}$, with $\xi_{it} \overset{iid}{\sim} N(0, 0.0459)$ and $\xi_{it} := \sqrt{0.7}\xi'_{it} + \xi''_{it}$ $\left(\xi'_{it} \perp \xi''_{it}\right)$. The firm observes ξ'_{it} when it chooses intermediate material and labour, with $\xi'_{it} \overset{iid}{\sim} N(0, 0.055)$. But both ξ''_{it} and ϵ_{it} are ex-post shocks, with $\xi''_{it} \overset{iid}{\sim} N(0, 0.0074)$ and $\epsilon_{it} \overset{iid}{\sim} N(0, 0.01)$.

There are three different productivity groups. Each group productivity has two components, a global factor, a^*_t, and a group specific factor, a^*_{gt}. For a given weight specification w, $\alpha^w_{g_it} = wa^*_{g_it} + (1-w)a^*_t$. These factors are new additions to Ackerberg, Caves, and Frazer (2015)'s DGP1 and I model them as AR(1) processes, $a^*_t = 0.7a^*_{t-1} + q_t$ with $q_t \overset{iid}{\sim} N(0, 0.01)$ and $a_{g_it} = \kappa_{g_i} + 0.7a_{g_it-1} + \tilde{q}_{g_it}$ $\left(\kappa_g \in \{-0.01, 0, 0.01\}\right)$ with $\tilde{q}_{g_it} \overset{iid}{\sim} N(0, \sigma^2_{\tilde{q}})$. The firm observes $\alpha^w_{g_it}$ when it chooses the intermediate material and labour inputs (the time point t'). $\sigma^2_{\tilde{q}}$ is 0.01 for Design 1 but varies for Design 2.

Conditional on K_{it}, L_{it}, and η_{it}, the firm's optimal choice of M_{it} is the solution to the quadratic problem $M_{it} + M^2_{it} - \exp(\eta_{it}) K^{0.4}_{it} L^{0.6}_{it} = 0$.

By Backward Induction, the firm chooses L_{it} to maximize $\mathbb{E}_{t'}[\exp\left(\epsilon_{it} + \alpha^w_{g_it} + \eta_{it}\right)]K^{0.4}_{it} L^{0.6}_{it} - W_{it}L_{it}$. Based on the firm's information set at choosing inputs,

$$B_{it} := \mathbb{E}_{t'}[\exp\left(\epsilon_{it} + \alpha^w_{g_it} + \eta_{it}\right)] = \exp\left(\alpha^w_{g_it} + 0.7\eta_{it-1} + \sqrt{0.7}\xi'_{it} + \frac{\sigma^2_\epsilon}{2} + 0.0037\right).$$

Then the labor's first order condition gives, $L_{it}(B_{it}, W_{it}, K_{it}) = \left(\frac{0.6B_{it}}{W_{it}}\right)^{\frac{5}{2}} K_{it} = \left(0.6B_{it}\exp\left(-\log(W_{it})\right)\right)^{\frac{5}{2}} K_{it}$.

Accounting for policy functions of M_{it} and L_{it}, the firm's investment problem at the end of period t (denoted as t'') is to choose the sequence of $\{I_{is}\}^{\infty}_{s=t}$ maximizing

$$\mathbb{E}_{t''}\left[\sum_{s=t}^{\infty}(0.95)^s\left(B_{is}\left(0.6B_{is}\exp\left(-\log(W_{is})\right)\right)^{\frac{3}{2}} K_{is} - \frac{\phi_i}{2}I^2_{is}\right)\right]$$

subjected to the intertemporal constraint $I_{is} = K_{is+1} - 0.8K_{is}$.

As in Van Biesebroeck (2007), the Euler equation for investment is

$$\phi I_{it} = 0.106\mathbb{E}[B_{it+1}^{\frac{5}{2}} \exp\left(-\frac{3}{2}\log(W_{it+1})\right)] + 0.76\phi\mathbb{E}_{t''}[I_{t+1}].$$

By substituting forward iterated versions,

$$I_{it} = \frac{0.106}{\phi}\mathbb{E}_{t''}[\sum_{s=0}^{\infty}(0.76)^s B_{it+1+s}^{\frac{5}{2}} \exp\left(-\frac{3}{2}\log(W_{it+1+s})\right)].$$

Using independence and model's assumptions, working out tedious algebra gives, (subjected to rounding error after the 5th decimal)

$$I_{it}(w, a_{g_{it}}, a_t, W_{it}, \eta_{it}, \phi) = \frac{0.106}{\phi}\sum_{s=0}^{\infty}(0.76)^s f_1(s, W_{it}) f_2(s, \eta_{it}) f_3(s, a_{g_i t}, a_t, \mu_{g_i}, w),$$

where

1. $f_1(s, W_{it}) := \exp\left(-\frac{3}{2}(0.3)^{s+1}\log(W_{it}) + 0.005(1-(0.09)^{s+1})\right)$,
2. $f_2(s, \eta_{it}) := \exp\left(\frac{7}{4}(0.7)^s\eta_{it} + \frac{\sigma_\epsilon^2}{2} + 0.07875(1-0.49^s) + 0.06675\right)$, and
3. $f_3(s, a_{g_i t}, a_t, \mu_{g_i}, w)$
   $= \exp\left(\frac{5}{2}\left((0.7)^{s+1}\left(wa_{g_i t} + (1-w)a_t\right)\right) + w\mu_{g_i}8.33(1-0.7^{s+1}) + 2.45(1-(0.49)^{s+1})(w^2\sigma_q^2 + (1-w)^2\sigma_q^2)\right)$.

A truncated investment function is used in the simulation. The truncated function sums up to the hundredth term instead of up to infinity. The simulation initializes the firm's capital stock at zero. The Monte Carlo simulation only uses data from after running the simulated sector for five periods. This "burn-in" phase is employed to minimize the effect from how I initialize capital.

In the Monte Carlo setup, the presented ratio is equal to the expression

$$\frac{T-2}{T} \min_{g,g':g\neq g'} \mathbb{E}\left[\frac{\Sigma_{t=1}^{T}(\alpha_{gt}-\alpha_{g't})^2}{\Sigma_{t=1}^{T}\epsilon_{it}^2}\right].$$

Its scaled reciprocal, $\max_{g,g':g\neq g'} \mathbb{E}\left[\frac{\Sigma_{t=1}^{T}\epsilon_{it}^2}{\Sigma_{t=1}^{T}(\alpha_{gt}-\alpha_{g't})^2}\right]$, can roughly bound the probability of misclassification. From the Jensen's Inequality, $1 \leqslant \mathbb{E}\left[\frac{\Sigma_{t=1}^{T}\epsilon_{it}^2}{\Sigma_{t=1}^{T}(\alpha_{gt}-\alpha_{g't})^2}\right]\mathbb{E}\left[\frac{\Sigma_{t=1}^{T}(\alpha_{gt}-\alpha_{g't})^2}{\Sigma_{t=1}^{T}\epsilon_{it}^2}\right]$ and, thus, the expression is roughly inversely related with its reciprocal.

Next, I use a heuristic argument to derive the probability bound. In contrast to the asymptotic proof, the bound is based on classification using the population parameters $(\theta^0, \beta^{0,K}, \alpha_{gt}^0)$. However, the bound shows how the probability of misclassification relates to easy-to-understand moment conditions.

The inequality

$$\sum_{t=1}^{T}(y_{it}-x_{it}'\theta^0-p(z_{it})'\beta^{0,K}-\alpha_{g't}^0)^2 < \sum_{t=1}^{T}(y_{it}-x_{it}'\theta^0-p(z_{it})'\beta^0-\alpha_{g_0^i t}^0)^2$$

leads to misclassifying the ith entity as a group g' member, i.e. $\hat{g}_i \neq g_i^0$. From substituting the model into the inequality, expanding out and sending $K \to \infty$, the above inequality implies this second inequality

$$\sum_{t}^{T}\left(\alpha_{g_i^0 t}^0-\alpha_{g't}^0\right)^2 \leqslant \left|\sum_{t=1}^{T}(2\epsilon_{it})\left(\alpha_{g_i^0 t}^0-\alpha_{g't}^0\right)\right|.$$

This second inequality holds only when $\sum_{t=1}^{T}\left(\alpha_{g_i^0 t}^0-\alpha_{g't}^0\right)^2 \leqslant \sum_{t=1}^{T}4\epsilon_{it}^2$. If otherwise, Cauchy-Schwartz inequality implies $\left|\sum_{t=1}^{T}(2\epsilon_{it})\left(\alpha_{g_i^0 t}^0-\alpha_{g't}^0\right)\right| < \sum_{t}^{T}\left(\alpha_{g_i^0 t}^0-\alpha_{g't}^0\right)^2$, an absurd conclusion. Let $\nu_{g,g'}^T := \sum_{t=1}^{T}4\epsilon_{it}^2 - \sum_{t=1}^{T}(\alpha_{gt}^0-\alpha_{g't}^0)^2$. Hence, whenever $\hat{g}_i \neq g_i^0$, there is at least one group g' with $\nu_{g_i^0,g'}^T > 0$.

This observation gives an upper bound on the probability of misclassification

$$
\begin{aligned}
\mathbb{P}\left(\hat{g}_i \neq g_i^0\right) &\leqslant \sum_{g: g \neq g_i^0} \mathbb{P}\left(\nu_{g, g_i^0}^T \geqslant 0\right) \\
&\leqslant \sum_{g: g \neq g_i^0} \mathbb{P}\left(1 \leqslant \frac{4 \sum_{t=1}^T \epsilon_{it}^2}{\sum_{t=1}^T \left(\alpha_{gt} - \alpha_{g_i^0 t}\right)^2}\right) \\
&\leqslant 4(G-1) \max_{g, g': g \neq g'} \mathbb{E}\left[\frac{\sum_{t=1}^T \epsilon_{it}^2}{\sum_{t=1}^T (\alpha_{gt} - \alpha_{g't})^2}\right]
\end{aligned}
$$

The first line comes from misclassification implying at least one ν^T is non-negative. The third line comes from the Markov's Inequality. Thus, a smaller reciprocal implies a smaller classification error in this context.

Here, I describe the cross-validation procedure used in the Monte Carlo simulation. In that section, I chose the polynomial order to non-parametrically estimate the semiparametric $_{m}$. However, the procedure can be applied to the general series estimator. Let the set of candidate series be $_{\{p^{K_1}(z),\ldots,p^{K_J}(z)\}}$. The cross-validated choice is the series $_{p^{K_j}}$ minimizing the predictive 10-fold mean-squared error of $_{y_{it}}$.

Algorithm 2: p^{K_j}'s Predictive 10-Fold Mean-Squared Error of y_{it}

Estimate $_{\hat{g}_i}$ with $_{p^{K_j}(z)}$ and $_{x_{it}}$ from the full sample;

Randomly partition the full sample's cross-sectional units into 10 subsamples,

$_{(\mathfrak{S}_1,\ldots,\mathfrak{S}_{10})}$, to be roughly equal in size;

for *k in 1:10* **do**

  Using $_{\hat{g}_i}$, estimate the coefficients of $_{p^{K_j}}$ and $_{x_{it}}$, and $_{\alpha_{gt}}$ off the cross-sectional units only in $_{\mathfrak{S}_k}$'s complement;

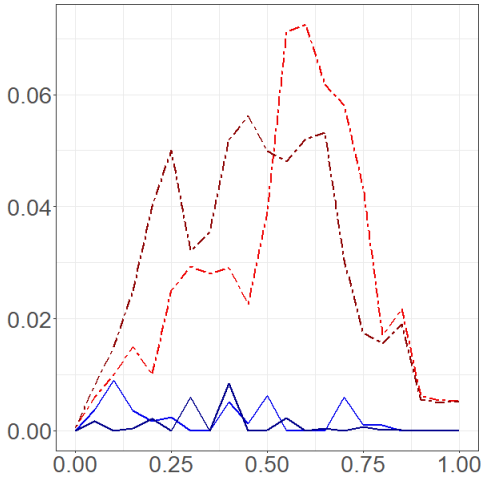
  Using $_{\hat{g}_i}$, calculate the mean-squared error of the previous step's estimated model on predicting $_{y_{it}}$ in the $_{\mathfrak{S}_k}$;

  Record this mean-squared error as the $_{p^{K_j}}$'s $_{\mathfrak{S}_k}$ predictive mean-squared error;

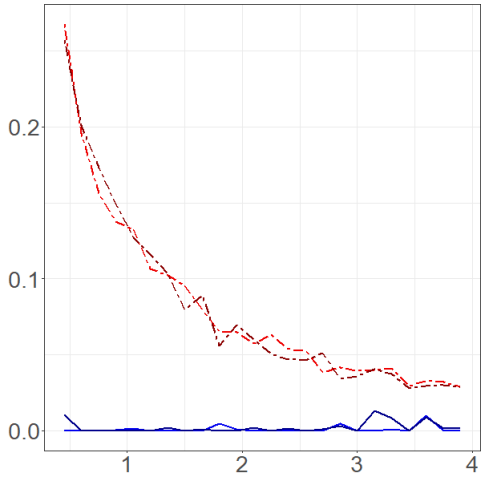
end

$_{p^{K_j}}$'s predictive 10-fold mean-squared error is its averaged 10 subsample predictive mean-squared error;

This is a very computation intensive process to do in simulation because the group memberships have to be re-estimated for every series choice. For cross-validation, I use only ten different group initialisations to compute the final group memberships. The small number of initialisations reduces my computation burden.



Design 1.

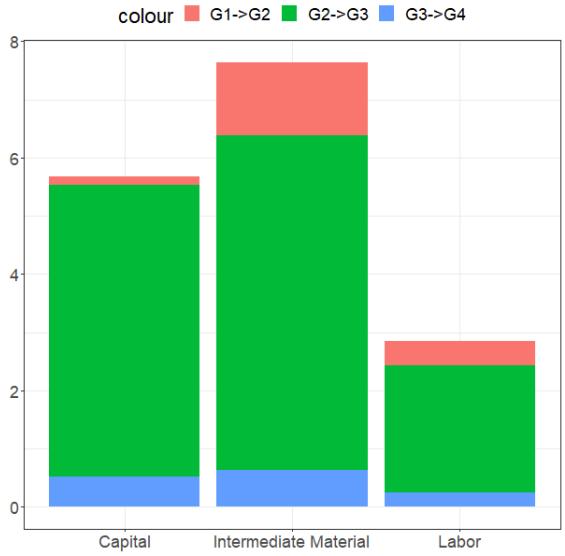


Design 2.

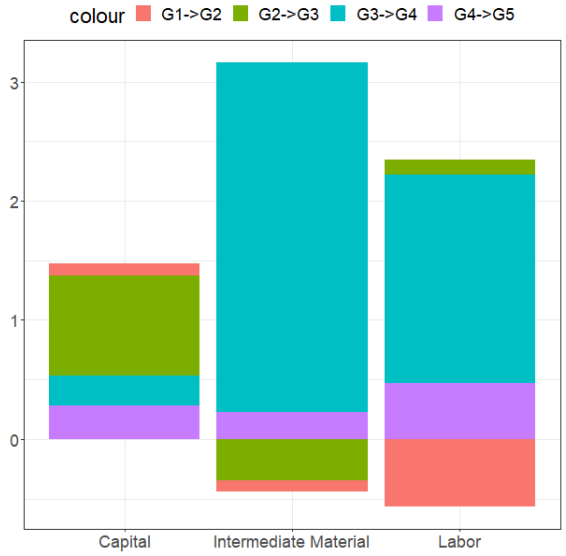
Y-axis: Average Classification Error | X-axis: w and $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon}$ for Design 1 and 2, respectively.

Solid/Blue line: $T = 20$. Dashed/Red line: $T = 5$. Blue/Red: $N_g = 100$ (Number of observations for each group). Dark Blue/Dark Red: $N_g = 300$.

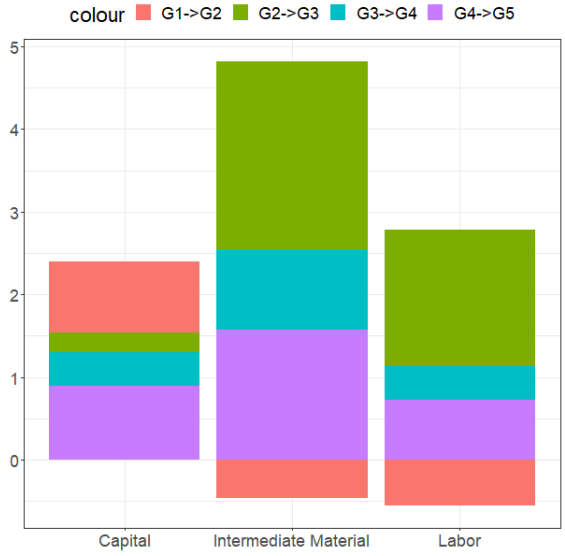
Empirical



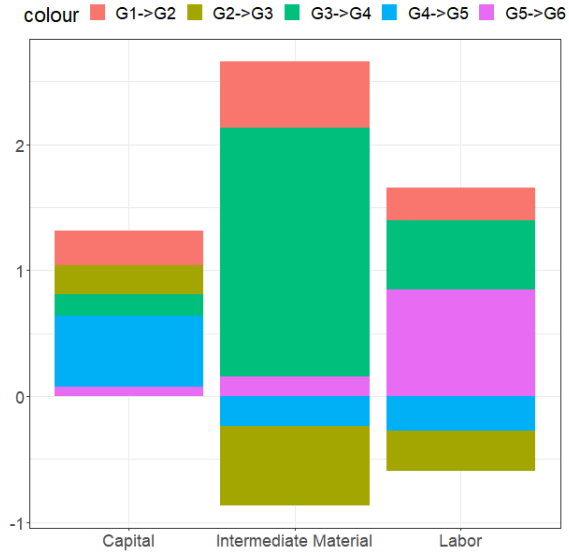
Food: Increment in mean level input choices



Textile: Increment in mean level input choices

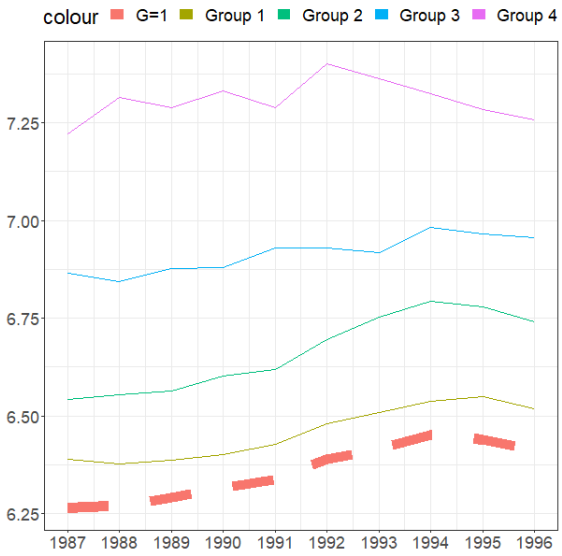


Metal: Increment in mean level input choices

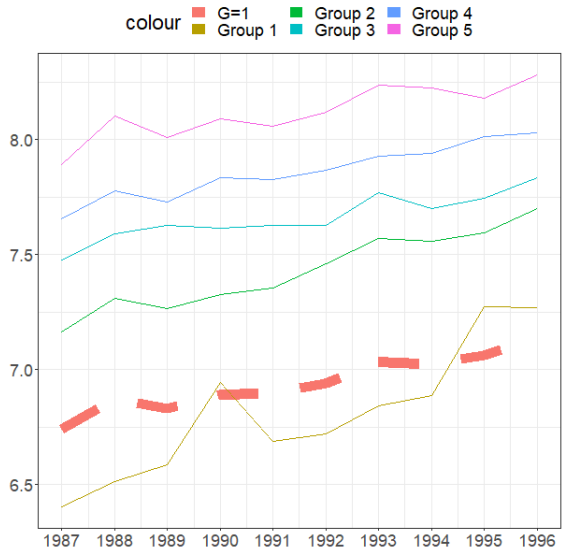


Wood: Increment in mean level input choices

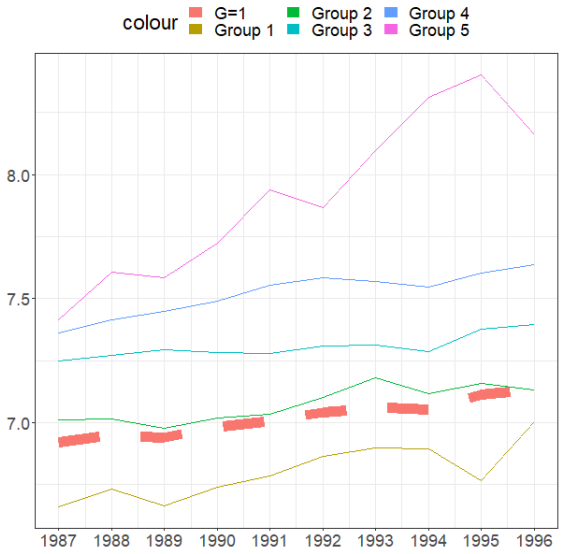
Description: Groups are ordered by their mean grouped gross productivity. Each stacked bar measures the factor increase of mean level input from one group to the next group. For example, the purple bar's size is the difference between Group 5's mean and Group 4's mean then divided by Group 4's mean.



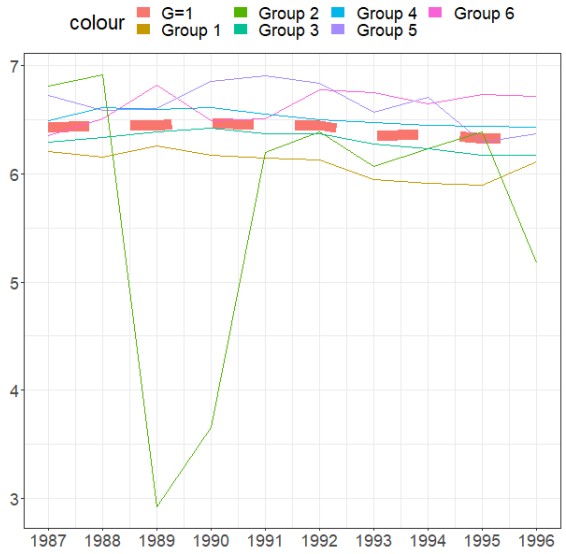
Food: $\hat{\alpha}_{gt}$ from 1987 to 1996.



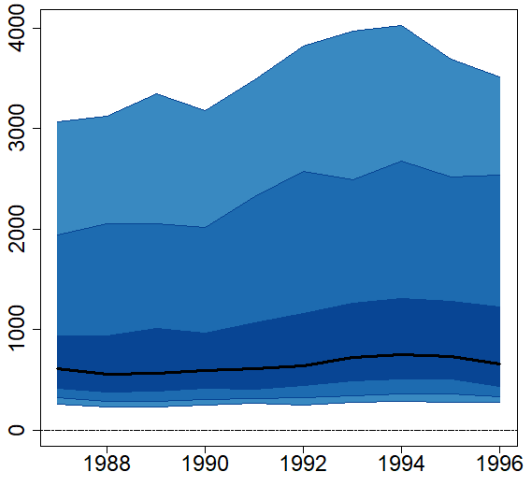
Metal: $\hat{\alpha}_{gt}$ from 1987 to 1996.



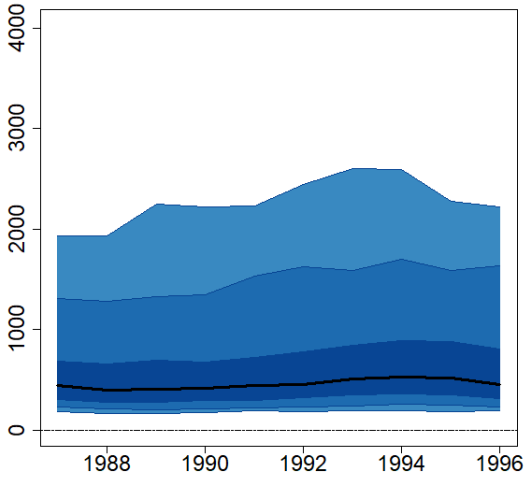
Textile: $\hat{\alpha}_{gt}$ from 1987 to 1996.



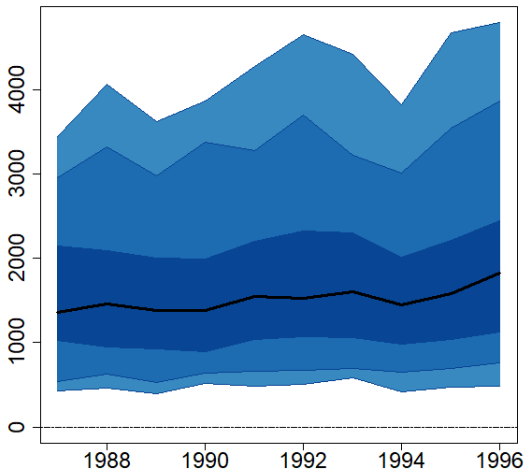
Wood: $\hat{\alpha}_{gt}$ from 1987 to 1996.



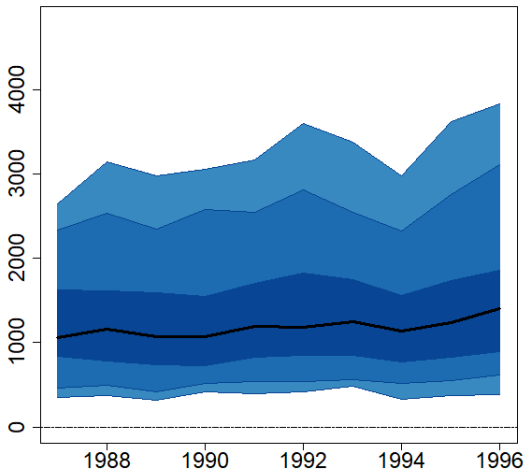
Food: $G > 1$ Productivity Fan Chart.



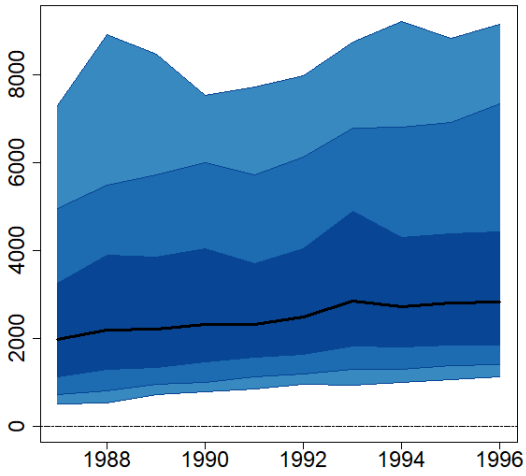
Food: $G = 1$ Productivity Fan Chart.



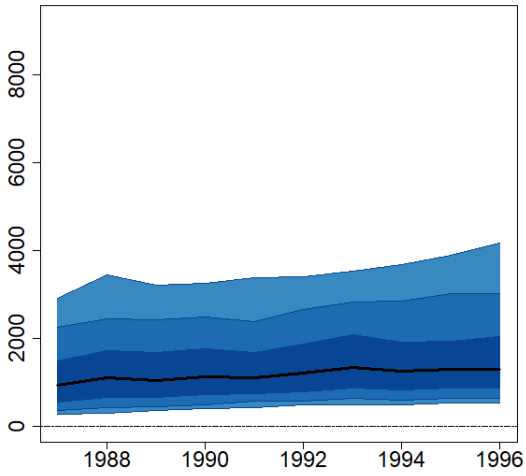
Textile: $G > 1$ Productivity Fan Chart.



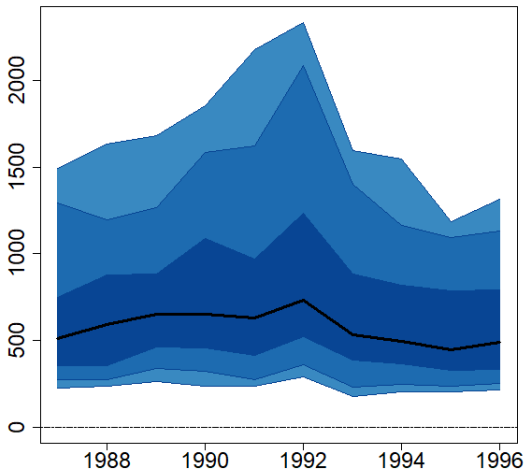
Textile: $G = 1$ Productivity Fan Chart.



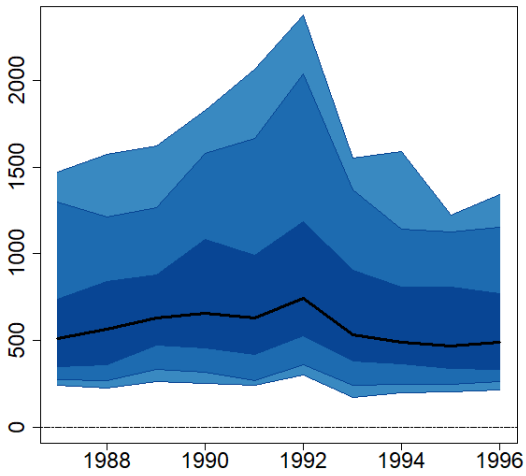
Metal: $G > 1$ Productivity Fan Chart.



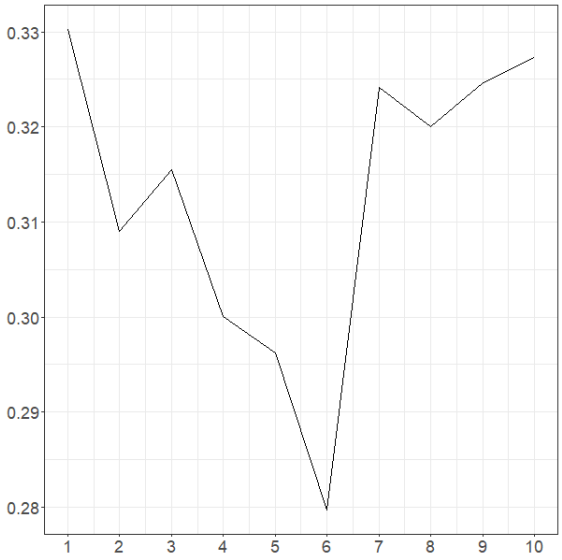
Metal: $G = 1$ Productivity Fan Chart.



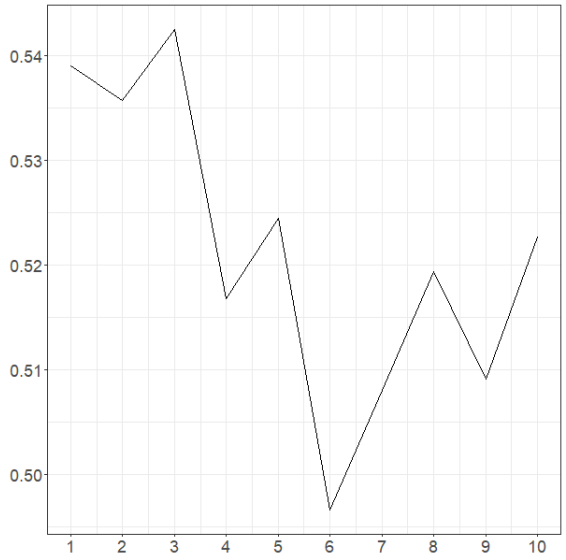
Wood: $G > 1$ Productivity Fan Chart.



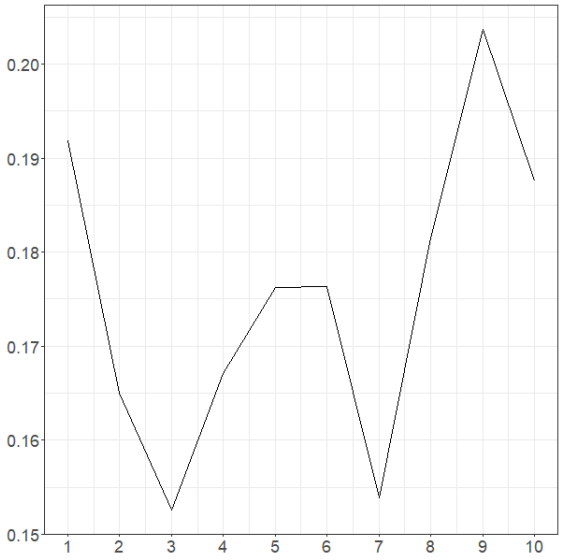
Wood: $G = 1$ Productivity Fan Chart.



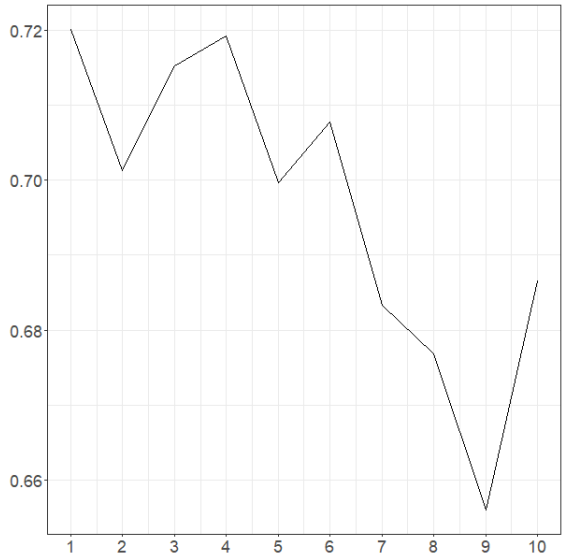
Food: Capital Coefficient over different G.



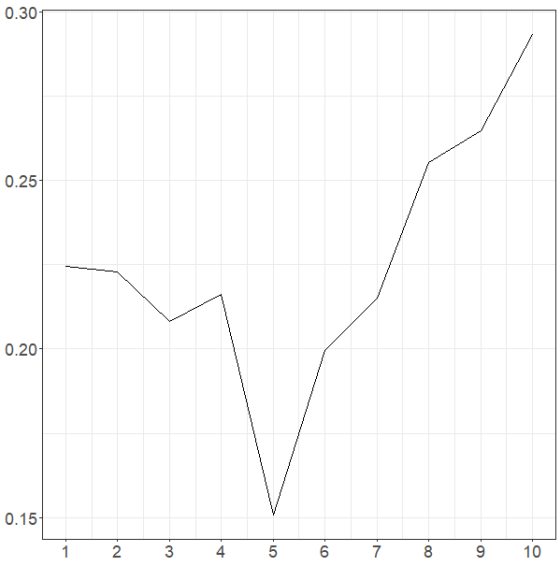
Food: Labor Coefficient over different G.



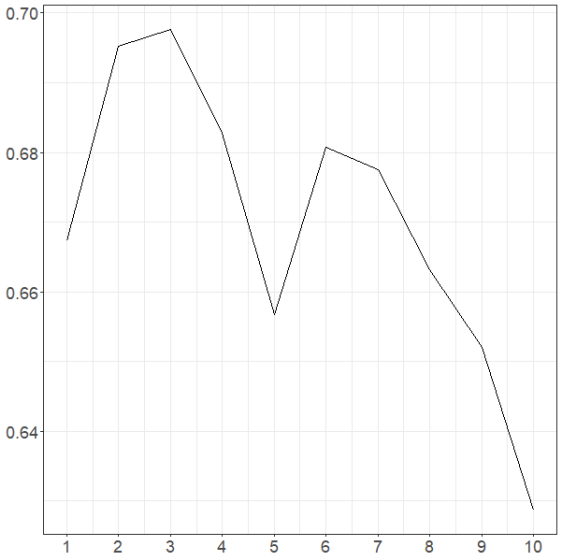
Textile: Capital Coefficient over different G.



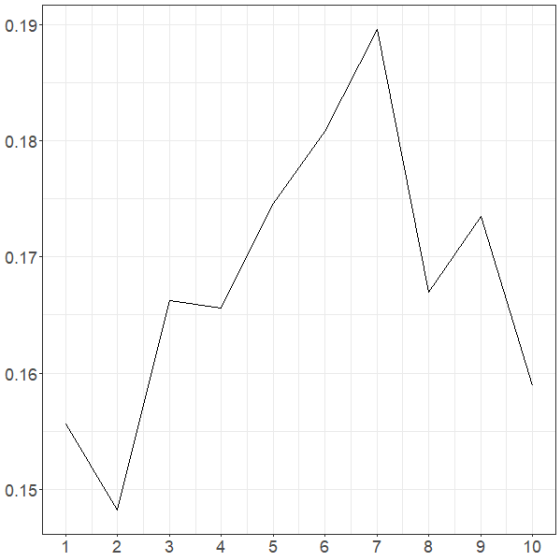
Textile: Labor Coefficient over different G.



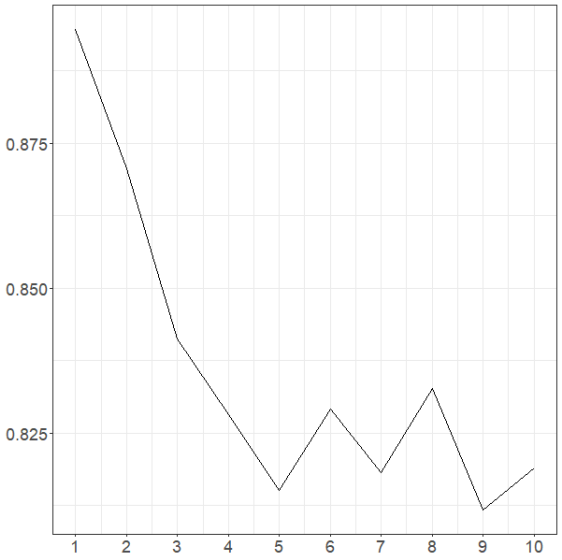
Metal: Capital Coefficient over different G.



Metal: Labor Coefficient over different G.



Wood: Capital Coefficient over different G.



Woof: Labor Coefficient over different G.